A Lifetime of Language Testing: An Interview with J. Charles Alderson

Tineke Brunfaut

Lancaster University, UK

Brunfaut, T. (2014). A lifetime of language testing: An interview with J. Charles Alderson. *Language Assessment Quarterly*, 11 (1), 103-119.

Professor J. Charles Alderson grew up in the town of Burnley, in the North-West of England and is still based in the North West, but in the ancient city of Lancaster. From Burnley to Lancaster, however, lies a journey and a career which took him all around the world to share his knowledge, skills, and experience in language testing, and to learn from others on language test development and research projects. Charles has worked with, and advised language testing teams, in countries as diverse as Austria, Brazil, the Baltic States, China, Ethiopia, Finland, Hungary, Hong Kong, Malaysia, Slovenia, Spain, Sri Lanka, Tanzania and the United Kingdom – to name just a few. He has been a consultant to, for example, the British Council, the Council of Europe (CoE), the European Commission, the International Civil Aviation Organization (ICAO), the UK's Department for International Development (DfID – formerly known as the Overseas Development Agency, or ODA) and the Programme for International Student Assessment (PISA) of the Organisation for Economic Co-operation and Development (OECD).

For over 30 years, however, Lancaster University in the UK has been his home institution, where he has taught courses on, for example, language assessment, language acquisition, curriculum design, research methodology, statistics, and applied linguistics. The list of post-graduate students supervised by Charles Alderson is

several pages long, and has resulted in a worldwide language testing alumni network. The co-authored handbook *Language Test Construction and Evaluation* (Alderson, Clapham and Wall, 1995) has introduced many a student and teacher to the basics of the field.

Charles was the co-editor of the successful *Cambridge Language Assessment Series*, which provides in-depth coverage of key areas in language testing. His own volume in this series, *Assessing Reading* (Alderson, 2000), is a written account of his expert insights into the construct of reading and its assessment. His preference for looking into reading can also be seen, for example, in his publications in journals such as *Reading in a Foreign Language*, *Language Assessment Quarterly*, and *Language Testing*, and in numerous book chapters, including the still frequently cited chapter 'Reading in a foreign language: A reading problem or a language problem?' (Alderson, 1984). Furthermore, Charles co-edited the journal *Language Testing* in the period 1997-2001, and guest-edited special issues on washback in 1996 and on assessment in Europe in 2005. He was the series editor of 'Into Europe: Prepare for Modern European Examinations' (see http://www.lancs.ac.uk/fass/projects/examreform/Pages/Exams.html).

Charles has undeniably played an important role in the founding of a global language testing community. In 1993, Charles became the first elected president of the International Language Testing Association (ILTA). In 2008, it was this same Association that acknowledged Charles' significant contributions to the field with the *UCLES/ILTA Lifetime Achievement Award* at the 30th Language Testing Research Colloquium in Hangzhou, China. His award talk was appropriately introduced with the title 'A lifetime of language testing', which later also became the title of a volume bringing together a wide-ranging collection of Charles' publications (Alderson,

2011a). In addition, Charles was the driving force behind the establishment of the European Association of Language Testing and Assessment (EALTA), the UK Language Testing Forum, and recently also the Second Language Acquisition and Testing in Europe network (SLATE).

In his free time, Charles has always been a fervent Munro climber (Scottish mountains over 3,000 feet in height). Unfortunately, two knee replacements have made him temporarily a Munro spectator, but he will soon be attempting the final 12 Munros, as he has already "bagged" 272 of them. Charles also enjoys travelling to remote corners of the world to get to know about other cultures and to admire nature in all its forms and beauty. Recent trips include Svalbard in the Arctic, Antarctica, and Myanmar.

On the occasion of his retirement, Charles spoke about his lengthy career in language testing with Dr Tineke Brunfaut, his colleague at Lancaster University, in a face-to-face interview on April 29, 2013. The interview was transcribed, and edited by Tineke (TB) and Charles (JCA) for publication in *Language Assessment Quarterly*.

TB: It is lovely to see you here in Lancaster! Despite having retired recently, you continue to spread the word about language testing and work on international research projects. Lancaster is unmistakably associated with you, and vice versa. But not everyone may know that your first appointment at Lancaster University in 1980 was preceded by a somewhat different and varied professional life. Could you talk about your early professional career and how this led to focussing on language testing?

JCA: Well, after I left Oxford University in 1967 with a degree in modern languages, I went to work in marketing for two years, in order to experience life outside education, at least for a short while. I eventually became product manager for Coleman's mustard, a job that had quite a bit to do with statistical trends, for example sales figures, which I found quite interesting, but the job itself was pretty boring after a while. So, after two years, I decided I should maybe go back to using my languages, as I had taken French, German and Spanish at A-level[i] and studied French and German at Oxford University. So I thought, "Well, I could either teach German in England or teach English in Germany." Since my English is better than my German, I got a job in Germany! I joined the recently founded Düsseldorf University in what was then the Federal Republic of Germany. This was 1969, and the very first task I was given was to design a placement test for incoming students. However, I didn't know what a placement test was. I didn't even know what testing was, really, and I hadn't done any language teaching yet…

Luckily, I shared an office with Dr Gerold Deffner, who quickly became a firm friend, and he told me about a book by Robert Lado. I managed to get hold of his book on language testing (Lado, 1961). So I read it, then sat down, and wrote some test items. The result was pretty embarrassing; a fairly awful test really ((laugh)). But it served to place students into different groups, which was, after all, its purpose. It was very grammar-vocabulary focused, but what I really enjoyed about this task was analysing the results afterwards. I learned about item analysis: item difficulty, item discrimination, and about reliability. To my surprise, some of the items were not bad (although others had quite a few problems). Nevertheless, the test served its purpose. So that was my first

"professional" contact with language testing, other than having been tested throughout my life as a student, of course.

After a couple of years working as an English lecturer at Düsseldorf University and learning on the job, I decided that what I should do was study for a further degree in Applied Linguistics, in order to be suitably qualified to teach. I was accepted into the Department of Applied Linguistics at Edinburgh University, where I had applied for an MPhil degree, not realising that this was a research degree! The first year of the MPhil was actually a Diploma in Applied Linguistics.

TB: Were you expecting a teaching degree?

JCA: I didn't really know what to expect, other than to learn about language teaching as well as language testing. When I arrived in Edinburgh I was told I would be there for two years, which I knew, but also that my thesis title was 'The psycho-sociolinguistics of language teaching and testing'.

TB: A very broad title! ((laugh))

JCA: Indeed! During the first year, I took a course in language testing as part of the Diploma and I enjoyed that very much. Towards the end of that first year, I realised that I would have to go away for a couple of years to make some money to pay for my studies and to support my family, before coming back to actually do the research. Before leaving, I talked to Professor Alan Davies, who had taught the testing course, and I explained that I would quite like to do research specifically in the area of language testing, rather than the very broad topic I had chosen before. He said, "Good, good, that's excellent. Why don't you look at the cloze test?" You need to know that this was in 1972, when John Oller was publishing on cloze tests (e.g. Oller 1972; Oller, Bowen, Dien and Mason 1972).

Alan made various suggestions for issues I could look into, so during the summer before I went abroad, I did more reading on language testing in general and on the cloze test in particular. I designed some cloze tests and then I went to live in Algeria for two years, working with the British Council, at the University of Algiers as an English lecturer.

TB: Using the cloze tests?

JCA: Yes, I used the cloze tests on my students in Algeria, and that became part of the data for my MPhil degree. After two years, I went back to Edinburgh and administered the same cloze tests to other overseas students around Edinburgh and in Newcastle, and I also administered the same tests to native speakers.

TB: Why exactly did you administer the cloze tests to both second language learners and native speakers?

JCA: I thought it would be interesting to see how native speakers performed. The cloze procedure had been used initially as a measure of readability rather than reading ability. Wilson Taylor used it in Japan with English native speakers, although he wasn't judging the native speakers at first; he was judging the texts (Taylor 1953, 1956). It struck me as slightly odd – if native speakers didn't get perfect scores, which they didn't, what was this test testing? Therefore, the aim of my thesis became to compare native and non-native speakers of English on the cloze test.

About half way through that second year in Edinburgh, which constituted the second year of my MPhil degree, Alan Davies suggested that I switch to a PhD. I remember replying, "I'm not good enough to do a PhD", but he said, "Of course you are". So I switched my registration to a PhD and did my PhD on the cloze test.

TB: In the book ‘A lifetime of language testing’ (Alderson, 2011a), you write that your research on cloze tests is often forgotten and ignored. Why would you like younger generations of language testers to know about that research? Why do you want them to read it?

JCA: I think there is a general problem in the field of applied linguistics which is that people don’t pay much attention to the history of the field and what’s been found out or what is reasonably well known. They either try to do more original research, or what they read tends to be the most recent publications. Also, people who are training to be applied linguists or language teachers are not typically reading research articles. But over the years I’ve seen many people asking questions about cloze tests and using them without knowing the research literature. That’s a big shame because the research is worth reading. In fact, Alderson (in press) will include one or two of my articles on cloze.

TB: Are there particular findings of that research which are still extremely relevant and important today?

JCA: They are relevant, because people do still use the cloze. The most important finding, I guess, is that the cloze technique is just that: a technique for producing cloze tests. Christine Klein-Braley coined the phrase, ‘A cloze is a cloze is a question’ (Klein-Braley, 1983). Well, it’s not. The cloze procedure is simply a procedure for deleting words from text in a pseudo-random fashion (removing every nth word from text, where n is a number anywhere between 5 and, say, 14). The result of such a pseudo-random deletion procedure is known as a cloze test, but the results of such a test are rather unpredictable. In other words, it is unclear exactly what a cloze test – any cloze test – is actually testing. Therefore, every

cloze test has to be validated; merely using the cloze technique to create a 'test' does not automatically make it a valid or reliable test.

I actually wrote up my PhD research after I left Edinburgh, in Mexico. By this time I had a young family and I had to work at least part-time to support my family. I did part-time teaching in Edinburgh, and then in 1977 I got a job with the British Council in Mexico City. I became director of the Research and Development Unit within the Modern Language Centre at the National Autonomous University of Mexico, where we developed many placement and achievement tests.

TB: Using cloze tests?

JCA: No, using multiple-choice techniques ((laugh))! At some point, Larry Selinker, who was the Director of the English Language Institute at the University of Michigan in the US, came to a conference in Mexico and, as a result of that visit and our conversations, I was offered the post of Director of Testing at the University in Michigan in Ann Arbor, and so I moved countries once more.

TB: On your work in Ann Arbor, I have heard you mention that an unpublished report which you wrote with Jane Anderson on the Michigan Placement Test (Alderson and Anderson, 1979) is one of the studies you are most pleased with, because it raises issues that are still current in language placement and achievement testing. Could you tell us a bit more about this research, since it is not widely accessible?

JCA: Well, in general terms, there are two basic approaches to placement testing. One is the achievement approach, the other is the proficiency approach. Either a placement test will be based upon the syllabus of the courses that students are being placed into, which is the achievement approach. Or, they will be based upon a measure of proficiency because students come from very different

backgrounds, different curricula and syllabuses, and therefore it is not terribly fair to put people, who have learned English in very different syllabuses, through one that they have not yet seen. The Michigan Placement Test followed the proficiency approach, and was used to place students into the English Language Institute (ELI) at the University of Michigan. In our study, Jane Anderson and I administered the placement test at the beginning of the term to newcomers and also to students who were already studying at the ELI, in order to compare those newcomers with those who were already in the ELI and were going up to the next level. What we discovered was that most of the people who were being promoted to the next level should not have been promoted, according to the results of the placement test. Now, it could partly have been a problem of inappropriate cut scores being used, but we showed that there was a clear tendency to push students up to the next level and then on to the next level and so on, regardless of their actual proficiency level, and this was creating heterogeneous classes. For instance, the students who were placed at level three in a six-level system on the basis of the placement test results were quite a lot better than the students who got into level three from the level below. That finding raised all sorts of issues about the achievement of students during the curriculum and indeed about the equity of the placement process.

TB: After having worked at Michigan for nine months, you returned to the UK in 1980. You started working at Lancaster University, about which you have implied in the past that it is the place where you got the label – if I can put it that way – ‘language tester’.

JCA: When I came to Lancaster, I joined the Institute for English Language Education (IELE) which did pre-sessional teaching of English for Academic

Purposes and Study Skills, as well as running in-service teacher training courses. Most of my work involved running short training courses of typically three months or ten weeks for teachers who were sent by the UK's Overseas Development Agency (ODA- see above) and the British Council to learn more about communicative language teaching. Within this teacher training work, I was expected to talk about testing, given my background. In fact, although I had only just arrived in Lancaster, Chris Candlin, who was the director of the IELE at the time, asked me to give a talk to a group of publishers on communicative testing. I said, "I don't know anything about communicative testing. I've only just arrived here". To which he replied, "Of course you do. You know about communicative teaching". So, I had to make up a story about communicative testing, and already Chris had me labelled as 'the tester'.

The British Council started to consult with me specifically on language testing and sent overseas visitors to see me as 'the tester'. I remember on one particular occasion, it was probably in 1980 or '81, there was a visitor who knocked on my door and said, "Doctor Alderson? Oh, you can't be. You're too young" ((laugh)). So that was part of 'the labelling process'. But to be fair, Caroline Clapham, who I knew from Edinburgh, came to Lancaster before me and had taught testing courses before I arrived. So there already was a language tester at Lancaster University, and together we taught language testing on the MA programmes in the Department of Linguistics and English Language (at the time called 'Linguistics and Modern English Language'), on behalf of the IELE.

TB: I have heard you say to students that as a language tester you have to be more than a tester. The Linguistics and English Language department at Lancaster, with a wide range of linguistics and applied linguistics areas represented, must

have been an interesting place to come to in this respect. Could you elaborate on interactions with other applied linguistic areas and other fields which you consider important for language testers?

JCA: One of the attractions of Lancaster to me was indeed that it had a lot of well-known applied linguists. But my first interest in testing, as I've mentioned, was actually the statistical side. It was the hard statistical evidence, if you like, about the quality of the test that fascinated me. So to be a tester you have got to have some interest in psychometrics. It doesn't necessarily have to be extremely deep; indeed, I've never really specialised in psychometrics as such. I usually quote the following story to my students: 'Statisticians are looked down upon by mathematicians, who consider statistics to be a misapplication of mathematics. Statisticians look down on social statisticians because they see it as watered-down statistics, and social scientists or statisticians look down on psychometricians because they're not interested in the underlying constructs which the statistics are supposed to help explain. And language testers are looked down on by everybody of course' ((laugh)). Nevertheless, I think it's very important for a language tester to know about language and not just statistics. Since my first training was in languages, all sorts of areas of applied linguistics have interested me in one way or another, from second language acquisition through reading theory through academic writing and so on and so forth. To my mind, you cannot be a test developer or a testing researcher, if you don't know about language constructs. That may seem fairly obvious, but it's not always the case that language testers pay attention to construct rather than statistics and psychometrics. But I think you've got to be balanced; you need to be interested in, and good at, both.

TB: Which areas in applied linguistics in particular should we interact with as language testers?

JCA: I don't think it matters actually, as long as it is related to language constructs. For example, through my work with cloze tests I got interested in reading in a second or foreign language, and I have stayed interested in it ever since, as well as, to some extent, listening. There is so much to know by studying reading - grammar, vocabulary, pragmatics – all those things. Any aspect of applied linguistics in my view can be relevant to language testing. Even critical discourse analysis can inform an approach to testing and assessment. For example, particularly at the higher levels of proficiency, looking at learners' ability to separate fact from opinion or to identify bias in texts or poor argumentation can be informed by work conducted in critical discourse analysis.

TB: Your interest in reading runs like a thread through your research and publications, and you've looked at it from many different angles. Of your earlier work the paper 'Reading in a foreign language: A reading problem or a language problem?' (Alderson, 1984) is probably one of the most well-known of your publications on reading. It is still often quoted, and has generated a lot of research. What do we know now about foreign language reading which we didn't know in 1984? What are the implications of our knowledge of the construct of reading for the testing of reading?

JCA: I got interested in the topic of foreign language reading during my time at the Modern Language Centre in Mexico, where the students were learning to read in English. That was the skill they had to master and which was tested. They needed to pass a test in order to be able to graduate, and a lot of the teaching in the centre was the teaching of reading, partly driven by the testing system. There was a very

common belief in Mexico in those days that Mexican students didn't read in Spanish and they were not particularly literate. They read cartoons, but they didn't read books and certainly not academic books. The belief was that English reading problems were not a language problem, but were due to the fact that students didn't read. It was thought that if you could teach students to read in their first language, it would transfer to the second language. That was a very strong belief, and it still is in many places. However, I asked myself, "Well, where's the evidence?" I didn't believe that people at university level didn't read in Spanish and I thought, "Let's test it". The 1984 article suggested ways of testing it, and basically raised a series of research questions and hypotheses. There still are researchers around who hold the view that first language reading does transfer to second language reading, even though the results show that, actually, language knowledge is more important than first language reading ability. Obviously, there has been a lot of research in first language reading, particularly for English, and researchers like Bill Grabe have been very influential in summarising and translating the findings from first language reading into a second language context. That has been very useful. Grabe has drawn our attention to a lot of the research that deals with cognitive aspects of reading in a first language (see, for example, Grabe 2009). This has been particularly helpful for the research project that I am currently engaged in on the diagnosis of reading problems, called DIALUKI (https://www.jyu.fi/hum/laitokset/solki/tutkimus/projektit/dialuki/en). In this research project we are looking at aspects of cognition as measured by different cognitive tests in both the first language and in the second language. We are examining how well these tests predict reading ability in both the first language

and in the second language. Our results show very clearly that the correlation between, for example, working memory tests or various other cognitive constructs and reading in the second language is much higher when those cognitive tests are in the second language than in the first language. In cognitive tests in Finnish and in English, given to Finnish-L1 learners of English, the predictions are much higher of those cognitive tests in English to reading performance in English than they are of cognitive tests in Finnish to reading in Finnish. That suggests to me that what is called the threshold hypothesis still holds, namely that foreign or second language reading is more of a language problem than a reading problem. What learners have to learn is the second language, through whatever means is appropriate, before any strengths they have in their first language can transfer.

TB: Is there a particular aspect of this research that has an impact on language testing?

JCA: Yes, diagnostic testing. A diagnostic test in particular has to be based upon a construct of what constitutes strengths and weaknesses, particularly weaknesses. You can argue that the strengths can compensate for the weaknesses, but we are typically more interested in weaknesses than in any compensatory factor, because we don't know what causes weaknesses, or strengths for that matter.

TB: What should current researchers be looking into as far as reading is concerned?

JCA: One area that is coming along in terms of research topics is the influence of the first language on the acquisition of the second language, so it's a second language acquisition (SLA) problem, if you like. It is interesting to look at reading from that perspective. It's the old contrastive analysis hypothesis (Lado,

1957), and we know that your first language will influence how easy or difficult it is to learn to read or to speak or listen in the second or foreign language.

TB: What could the role be of language testers in this type of research?

JCA: In my view, language testers and SLA researchers should know more about each other's field because testers know about measurement, know about task design, know about the results of poor task design, and SLA researchers can learn from us, just as we can learn from them what sorts of tasks might be useful, what sorts of things they find to be important in the learning process. Unfortunately for me, most SLA research is done on speaking, not on reading ((laughs)).

TB: As you suggest, language testers are quite thorough in their research methodology. Apart from looking into the construct itself, your reading research has also partly explored the methodological side of investigating reading. For example, you have looked into the use of judgements in research on (testing) reading (Alderson, 1993). In more recent research with our PhD student Gareth McCray, we have looked into the use of expert judges when investigating reading (Alderson, Brunfaut, McCray and Nieminen, 2012). Bringing all this research together, it seems that one of the methodological conclusions is that working with judges is quite hard because they don't necessarily agree with one another. Does this have implications for language testing research? How should we use expert judgements to contribute to our research design or findings, or should we stop relying on this method?

JCA: I don't think we can stop using judges; all professions use judgement, inevitably. Some judgments are good and some are bad, and some people are novices rather than experts, but the answer is triangulation. Alongside your

judgement methods you can use, for example, eye tracking to see what people are doing with their eyes. You can use verbal protocol analysis to see what learners say in think-alouds about the process. You can use statistics as well to throw light upon the results of tests or tasks. It is about a balance between those different approaches. Of course, most test developers don't have the luxury of using all those research techniques when developing tests. There are so many practical pressures on producing tests. It is often up to testing researchers to do research such as this.

TB: What is the unique contribution of the judgements in such mixed-methods design?

JCA: It is the view of construct – at whatever grain size you are making judgements, whether it is a broad judgement, for example, about whether something is a grammar item or a lexical item, or whether it is a judgement about what particular subskills are being tapped by a given reading test item. Those are all judgements. And matching item writer intentions against test results, that is important. Looking at specifications and seeing how the items themselves reflect those specifications and how performances reflect what you believe you have tested.

TB: A related issue is the question 'who is an expert' – a difficult concept to define.

JCA: An expert is an expert is an expert… ((laugh)). I've just published an article with one of our MA students (Alderson and Kremmel, 2013) on precisely this issue, looking at the judgements of so-called expert applied linguists. One of the reviewers of an earlier version of the paper asserted that the best people to make judgements about test content were item writers. That's an opinion. It's a judgement. It needs to be falsified or verified. In my experience there are good

item writers and bad item writers, and good item writers are often good at producing certain sorts of tests, the ones they have most experience of, and not necessarily other sorts of tests. It is often argued that, besides selecting experts, you should train the judges, but in the sort of research I've done I'm not interested in cloning people into making the same decision as everybody else. I want to know whether somebody who knows about language will agree with somebody else who knows about language, not because he has been cloned, but because they share an understanding of the construct.

TB: Another methodological area, for which it is fair to say that you, together with our colleague Dianne Wall, have moved the field forward, is washback research. Your 1993 article proposing a series of washback hypotheses is seminal (Alderson and Wall, 1993). What brought you to looking into this topic originally?

JCA: Rather similarly to the Mexican experience (see above), it was a set of naive assertions about testing and the impact of testing on learning or on teaching, or on learners and on teachers. It seemed to me that these assertions needed verification, which is why the title of that article is 'Does Washback Exist?' (Alderson and Wall, 1993). I read quite a few articles from the 1960s and '70s before that article was published, which contained a lot of assertions but no evidence. So Dianne Wall and I were interested in exploring that, both in our teacher training activities in the IELE and in the work we were doing in Sri Lanka at the time (Wall and Alderson, 1993; Wall, 2006). We were commissioned to develop a new O-level exam, a secondary school-leaving exam in Sri Lanka. The belief was out there, and indeed I have written to this effect, that if the test was a good test, the impact of the test would be good; only a bad

test has bad washback. In short, you can have positive washback and negative washback. We conducted research to prove to the ODA that the Sri Lankan tests were having a positive influence on teaching. But we were shocked by the results. A team of researchers observed teaching in schools in Sri Lanka, and we were surprised to discover that, although the teachers were indeed using materials that they believed were similar to exam materials, how they were teaching didn't change at all. We looked at ordinary classes and exam classes, and classes close to the exam and further away from the exam, but the teaching methodology didn't change. It was pretty poor and certainly didn't reflect what teachers were being taught in in-service courses on teaching reading. But the exam content was visible in the classroom: certain sorts of texts were being used and the teachers were concentrating on the language in the text.

Dianne's work is actually the most important in this area, because she later did a longitudinal study of the new TOEFL test with Tania Horak which shows that there can be good washback (Wall and Horak, 2006, 2008, 2011). But even with the TOEFL iBT, different teachers will teach towards it in different ways, despite the test preparation materials that exist. That is similar to the findings about the washback of the old TOEFL (Alderson and Hamp-Lyons, 1996).

The question is whether positive or negative washback is an issue of test design, which is what Sam Messick asserted (Messick, 1996), or it is a problem of the teaching, rather than a problem of testing. And then there is the question: can a good test be taught badly? Yes. Can a bad test be taught well? I don't know. I don't think anybody has looked at it, but it's an interesting question. You would, of course, have to decide what you mean by good and bad. But it is clear

that it is the perceived consequences of the testing that cause behaviour to change or not to change.

An interesting question is 'Can diagnostic tests have washback?' Or, do they have washback? Negative washback. I don't know the answer yet.

TB: That is something to explore further. But with regard to diagnostic testing, I guess the first question is 'Is there a truly diagnostic test?'

JCA: There are lots of tests that claim to be diagnostic… In the DIALUKI project, we are working on the answer to the question 'What should a diagnostic test look like?', for example, what features would a diagnostic test have? The argument is that diagnostic tests don't have consequences as they are low stakes, or have no stakes. That is what I've always argued about DIALANG, for example (Alderson 2005; http://www.lancaster.ac.uk/researchenterprise/dialang/about.htm). However, the fact is that DIALANG is more often used for placement than for diagnosis, and placement brings consequences with it. Interestingly, to my knowledge, nobody has investigated whether DIALANG has negative washback.

TB: Washback research seems to fit in particularly well with more recent views of validity. As you said, in the end it's all about consequences and usage of tests. For that reason, should washback be given an even more important place in research?

JCA: I don't see washback as being part of validity – I don't agree with the notion of consequential validity – for the reason I've hinted at, namely that I don't think washback is necessarily caused by the test. It is caused by the use or misuse of the test. Of course, we know that the use of the test relates to validity, but is a knife the cause of murder? What is important is who uses the test and how they use it, and whether it is used for high or low stakes. We know that TOEFL or

IELTS, for example, aren't bad tests. There are many studies of their validity, but would bad washback threaten their validity? I don't know; I don't think so. Sam Messick (1996) said "Seek washback by design", that it is the test developers' problem, but that doesn't guarantee you'll get washback or positive washback.

TB: So do you see washback research as more independent?

JCA: It needs to be broader. It needs to look at intentions and aims, for example of educational politicians or the people who are training the teachers. One thing I've talked about quite a lot is that many people see test preparation as cheating, but the best thing you can do for your students is to help them pass an exam. However, do we give enough attention in teacher training, for example, to how best to prepare for an exam? No.

TB: This area touches on a lot of political issues, doesn't it, which is not something you've been particularly shy of discussing throughout your career? You don't seem to have steered away from or been afraid of facing sensitive, political issues (see for example, Alderson, 2009).

JCA: It is partly because I tend to speak my mind ((laugh)). I am 'a northerner' (born and raised in the North of England), so that's part of my nature; that's what many northerners are like. But it is more that I am a researcher and I want to understand what makes people do what they do. It is clear that politics is an important part of that. I have quite a lot of experience of educational projects, which are often political – with a small p. I have seen political aims such as poverty alleviation or selling textbooks. If you ignore the politics of, for example, educational testing, you are ignoring test purpose, and that's validity. So, you've got to come to terms with it.

TB: One context in which I assume you have had to work with politicians, and so also within politics, is your work in Europe, for example the Hungarian exam reform project (http://www.lancs.ac.uk/fass/projects/examreform/Pages/Projects.html) and the Austrian secondary school-leaving exam project (http://www.uibk.ac.at/srp/). What were your experiences there?

JCA: The Hungarian project is the one that is closest to my career, because I spent many years working on it and was deeply involved in the politics of change in Hungary, from the political openings that happened in 1990-91 through to the political tensions within institutions, which were partly the result of personal politics of people who didn't like the British Council being involved in supporting educational reform. I saw similar political problems, with a capital P, in the work in Sri Lanka (see above), for example racial problems. The exam reformers in Sri Lanka were accused of being agents of the British. When working on practical, real-world exam reform projects you see politics up close. I have seen resistance to change in projects in Austria, the Baltics, Hungary. If you are working in testing, you just can't avoid politics. Whether you want to write about it or research it is another matter, but it is there.

TB: What is the significance of these reform projects?

JCA: Well, one aim, of course, is positive washback, to steer the teaching in some way or another. The Council of Europe has political aims to encourage use of the Common European Framework of Reference (CEFR; Council of Europe 2001), and DIALANG contributed enormously, but innocently, to that aim (Alderson, 2005). Some people see that as interfering; some roundly condemn encouraging use of the CEFR.

It is clear in Europe that the CEFR has been most successful in the area of influencing language test development. It has been less successful in influencing teacher training or curriculum design. But it is in assessment where it has had most traction, which means that the politics of innovation and resistance to change need to be researched and understood.

TB: Do you have any tips on how to work with policy makers who may not necessarily have a language education background?

JCA: I wish! ((laugh)) Well, of course the buzzword these days is assessment literacy. Real politicians, rather than small p politicians, need to realise that assessment is one aspect of education that has to be taken seriously. Assessment should not be used as a hammer 'to beat' teachers or learners; that is misuse of tests, in my view. But politicians are necessarily short-term in their thinking. They worry about the next election and so they want their innovations to happen within the length of a parliament, unless you can get – as seems to be happening in Austria – both sides of a political divide to agree that change is needed, that it needs to take a certain direction and should include assessment.

People talk a lot about assessment literacy, but I don't really know how they go about educating politicians. In Austria, I have managed to talk to politicians and senior civil servants, who were keen to listen. People like Diane Schmitt, who is chair of BALEAP (a forum for EAP professionals; http://www.baleap.org.uk), has had conversations at least with civil servants about assessment and in particular the UK Border Agency's requirements of test-based evidence of language proficiency before visas can be issued. In Hong Kong, Lyle Bachman did a survey for the Hong Kong government (Bachman, 2010), in which he was highly critical of the use of IELTS as a graduation

requirement in Hong Kong tertiary institutions. The educational politicians listened and as a consequence cancelled the requirement to take IELTS before students could graduate from university. The institutions themselves now have to develop their own language proficiency tests as one of their graduation requirements, and that is no bad thing. Sometimes talking to politicians works.

TB: In your work in Europe, you were also involved in setting up the European Association for Language Testing and Assessment (EALTA). How exactly did EALTA come about? What were the original motives to develop this organisation? What gap did it fill?

JCA: Good question. It came about originally as a result of my experience with the International Language Testing Association (ILTA) where I tried to persuade people in Europe in particular to join ILTA. However, people in countries like Hungary, which is a fairly poor country, just could not afford ILTA's membership fees, which I saw as a problem. Secondly, although I was the first elected president of ILTA, ILTA was very much dominated by arguments about procedures and rules and regulations. It wasn't really getting on with creating a professional organisation and talking about principles of language testing. Fortunately, that has now changed, but it took a long time. So, when we were working on DIALANG, people in the European Commission, who were very much in favour of improving language tests, saw the need for teacher training to encourage test reform. It was suggested to me that it would be good if we were to work towards an association for language testers in Europe. I pointed out that there already was a European association, the Association of Language Testers in Europe (ALTE), but that that was an association of examination bodies only, and did not meet the needs of ordinary language teachers. So the European

Commission made resources available to set up an association specifically aimed at language teachers, which we called the European Network for Language Testing and Assessment (ENLTA), which later became EALTA. The first EALTA conference was held in 2004 in Kransjka Gora in Slovenia, and since then the association has gone from strength to strength.

TB: What role do you think EALTA currently plays?

JCA: Well, that's a question I and others are asking ourselves right now. EALTA developed guidelines for good practice, for example, which are now available in thirty-four languages. But the main activity of EALTA seems to have been to hold conferences and meetings of special interest groups, whereas I think EALTA now needs to broaden its base. It needs to reach out more to educational politicians and engage more vigorously in teacher training and advocacy for the improvement of language testing and assessment. EALTA has run a few summer schools, which is good, but it is currently an association which isn't really growing: it is not reaching out as much as it should. The association should consult much more with people – and especially younger people - about what sorts of things *they* feel the association should be doing and EALTA should then ensure that the ideas are implemented.

TB: You have mentioned ALTE, EALTA, ILTA, and there are also regional language testing organisations in Japan, and in Australia and New Zealand, for instance. In your view, what is the responsibility of such organisations?

JCA: Basically, spreading the word that language assessment is a profession and it needs to be taken seriously, that ad hoc assessment procedures are not a good thing, and that we need to guarantee the quality of tests and assessment procedures. Making a bridge between researchers and teachers, but also non-

teachers such as educational politicians, civil servants, all the decision-makers, starting with the Secretary of Education. Engaging teachers in creating tests and working with tests and learning about how things can be improved. For example, the Hungarian exam reform project (see above), had quite an impact on teachers. Although the tests that the Hungarian exam reform project produced were never properly implemented, there is now much more awareness in Hungary about issues of test quality and test development than there was before, which must be a good thing.

TB: An example of where a testing organisation, in this case ILTA, has reached out is the area of aviation English testing. You have personally been involved in this, drawing people's attention to the lack of evidence for the quality of many tests of aviation English, and working with the International Civil Aviation Organization (ICAO) to develop a system of test endorsement (known as the ICAO-Aviation English Language Test Endorsement scheme). In our field, there are recurring discussions on the nature and the extent of test developers' responsibilities, with different views on the topic being held. What is your view on this?

JCA: One of the reasons I got involved in aviation English testing was that this is an obvious case where the lack of good tests is potentially life-threatening (Alderson, 2010, 2011b). Lancaster University was invited to advise on the development of a test of English for Air Traffic Controllers (known as ELPAC – see http://www.eurocontrol.int/elpac-tests) by Eurocontrol (the European Organisation for the Safety of Air Navigation). The need for good tests could not be more obvious than in aviation, and the development of the aviation tests was guided by the EALTA Guidelines for Good Practice. There was some opposition within EALTA and ILTA to the idea of "policing" the quality of language tests,

but common sense eventually prevailed and the Aviation English Language Test Endorsement scheme finally saw the light of day a couple of years ago. Monitoring the quality of language tests needs constant vigilance, however, and we can never be complacent about quality control as one of the most important responsibilities of a language testers' association.

In terms of such responsibilities, I think you have to start with high-stakes testing, like aviation English testing, testing for university admission, for citizenship and immigration, the testing of doctors and nurses, and arguably also the testing of solicitors and so on through the professions. I suppose that compromises in quality are inevitable where the stakes are not so high, but they are not acceptable in high-stakes settings.

TB: You've used simulation videos of aviation accidents in introductory presentations to students, to illustrate how high the stakes can be and what the significance of language testing is. During your career, several generations of students have gone through your hands, for example on the online Masters in Language Testing you set up with colleagues at Lancaster. There is the annual Language Testing summer school at Lancaster, the training courses you have given all around the globe, and you have also pleaded for materials to be made freely available (for example, those of the Hungarian project on http://www.lancs.ac.uk/fass/projects/examreform/). The type of assessment literacy work you have done in this manner is perhaps less visible to the academic community, as this typically goes unpublished.

Based on your extensive experience, what do you consider to be key aspects of assessment literacy training?

JCA: Awareness is a big thing. To see that there is a problem is important, and that is why the aviation accident video I show is rather useful. But the core of assessment literacy training is the expansion of the concepts of validity and reliability and, to some extent, washback. What they mean, how to establish criteria for knowing something is valid or reliable, or more valid and more reliable, and has positive impact. That is basically what is needed. And equally important is enthusing people to get involved in something that is a challenge, but that can be enjoyable and satisfying: designing a test, figuring out what works and doesn't work, asking yourself why it doesn't work and what you can do to change it, solving problems. Part of one's enthusiasm comes from wanting to motivate people to take action, but also encouraging people to go off and do their own thing and helping them with it, particularly with PhD students or on assessment projects.

TB: The list of PhD students you supervised is long, including well-known testers such as Gary Buck, Caroline Clapham, Glen Fulcher, Jo Lewkowicz, and younger generations with Jay Banerjee, Spiros Papageorgiou, Alistair Van Moere – to name just a few. In 2009 this work was recognised by an award for Excellence in Doctoral Supervision at Lancaster University. What has been your approach to supervision? What is the secret?

JCA: Ha! Living a long time ((laugh)). Well, despite being a blunt speaker, I try to encourage students rather than to discourage them. It sounds trite, but you need to have patience. You need to pay attention to detail. You need to see your students as friends in one way or another. It's almost being a father figure, especially for those who come from abroad. But at the same time I also set students challenges that I think they can overcome and am fairly blunt when

they've got problems in their work. Obviously, you then need to put them on what you consider to be the right path. It's just common sense really.

TB: You officially retired from supervision, teaching and your post at Lancaster University towards the end of 2012, but you are not sitting still. You are still engaged in research on reading and on diagnosis. What makes you want to continue this line of research?

JCA: I think it is an under-researched area. I am attracted by the challenge of addressing the problem of what diagnosis is and how we can do it better. It is often claimed that you can produce diagnostic information from a proficiency test or from an achievement test, but I think that's unhelpful. It doesn't tell you what constructs you should include in a test specifically for diagnostic purposes. It is not at all clear that the cognitive diagnostic models we currently have are actually helpful in diagnosing strengths and weaknesses to help students to learn better, quicker, more thoroughly. That is the challenge. DIALANG has clearly been successful at one level, but it is unsatisfactory at other levels: a) in some of the languages involved in the programme, there aren't very many items, and b) the theory behind the test is fairly traditional. DIALANG was a test development project; it wasn't a research project.

TB: What is needed to move the area of diagnosis forward?

JCA: More and more work of the sort that we are doing in the DIALUKI project on diagnosing reading and writing in a second or foreign language (https://www.jyu.fi/hum/laitokset/solki/tutkimus/projektit/dialuki/en). It involves looking at all the different components that might affect somebody's weakness and strength in language learning. That means working more with SLA researchers. I think it is a pity that we are often pigeonholed as 'mere' testers,

and overworked as test developers rather than as test researchers. It would be good if we could get more people in SLA interested in co-researching and talking about constructs and how to measure them and how not to measure them.

TB: So it is back to constructs and research methodology – where we started this interview.

JCA: Absolutely.

TB: Do you have anything else on your wish list for future research?

JCA: It would be nice to see a lot more work, similar to what we've done in reading, on listening. That's an obvious important area. It would also be nice to see more work published from a diagnostic perspective in writing and speaking. It would be… where do you start? It would be nice to get a better handle on what the contribution of language is to the skills of language use. Which aspects of language should we focus on? Formulaic sequences? Language pragmatics?

TB: And do you have anything on your wish list for the practical side of language testing, for exam boards and test developers? What should they focus on in future?

JCA: In general terms, I think they should be doing more research into the quality of their instruments and into innovation, for example, looking at how to use new media and digital technology to deliver tests and assessment procedures. I would like to see exam boards being more open about what they do and how well they do it, being much more receptive to criticism of their work and more willing to change and innovate. I think we are seeing that with some exam boards, but they tend not to have many resources. I would like to see exam boards be less defensive about what they do, and do more research within the exam board and publish the results, be less defensive. It is hard work, though, being an exam

board. It is just grinding out items again and again; very repetitive and boring. Nine months in Michigan in 1979 was enough for me ((laugh)).

What I like about testing in general is the variety of things you can get involved in, different challenges. It is, after all, a very important area within the field of language learning and teaching!

TB: And on that note, I'd like to thank you very much for this interview, and wish you a productive retirement, but also a relaxing one!

References

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson and A. H. Urquhart (eds.), *Reading in a Foreign Language* (pp. 1-24). London: Longman.

Alderson, J. C. (1993). Judgements in language testing. In C. Chappelle and D. Douglas (eds.), *A new decade of language testing research* (pp. 46-57). Washington DC: TESOL.

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

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing*, 27(1), 51-72.

Alderson, J. C. (2011a). *A lifetime of language testing*. Shanghai: Shanghai Foreign Language Education Press.

Alderson, J. C. (2011b). The politics of aviation English testing. *Language Assessment Quarterly*, 8(4), 386-403.

Alderson, J. C. (in press). *[TBC]*. Shanghai: Shanghai Jiao Tong University Press.

Alderson, J. C., & Anderson, J. (1979). *Placement testing and achievement*. Internal Report, English Language Institute, University of Michigan

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test 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., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing* 0265532213489568, first published on July 2, 2013.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(1), 115-129.

Alderson, J. C., Brunfaut, T., McCray, G., & Nieminen, L. (2012). *Component-Skills Approach to L2 Reading: Findings, Challenges, and Innovations*. Paper presented at AAAL, Boston, USA.

Bachman, L. F. (2010). *Language Enhancement in Hong Kong Universities: Some Observations and Recommendations*. Consultant's Report to the University Grants Committee.

Council of Europe. (2001). *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* New York: Cambridge University Press.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. W. Oller, Snr (ed.), *Issues in language testing research* (pp. 218-228). Rowley, MA: Newbury House.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press: Ann Arbor.

Lado, R. (1961). *Language testing. The construction and use of foreign language tests: a teacher's book*. New York: McGraw-Hill.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.

Oller, J. W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modem Language Journal*, 56(3), 151-158.

Oller, J. W. Jr., Bowen, D., Dien. T. T., & Mason, V. (1972). Cloze tests in English, Thai and Vietnamese. *Language Learning*, 22(1), 1-15.

Taylor, W. L. (1956). Recent developments in the use of “cloze procedure”. *Journalism Quarterly,* 33(1), 42-99.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.

Wall, D. (2006). *The impact of high-stakes examination on classroom teaching.* Cambridge: Cambridge University Press

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, 10(1), 41-69.

Wall, D., & Horak, T. (2006). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, The baseline study. *ETS Research Report*, RR-06-18 TOEFL-MS-34. Princeton, NJ: Educational Testing Service.

Wall, D., & Horak, T. (2008). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, Coping with change. *ETS Research Report*, RR-08-37 TOEFLiBT-05. Princeton, NJ: Educational Testing Service.

Wall, D., & Horak, T. (2010) The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, The role of the coursebook. Phase 4, Describing change. *ETS Research Report*, RR-11-41 TOEFLiBT-17. Princeton, NJ: Educational Testing Service.

[i] The term ‘A-level’ refers to the General Certificate of Education Advanced Level, a secondary school qualification in England, Wales and Northern Ireland.