LTJ 28 1

[Start of recorded material]

Interviewer: From the University of Leicester in the United Kingdom. This is Glenn Fulcher with another issue of Language Testing Bytes.

In Issue 28(1) of *Language Testing* for 20/11, Khaled Barkaoui, who is Assistant Professor in the Faculty of Education at York University in Canada, has a paper on the use of think-aloud protocols in researching rate of behaviour. Now, think-aloud protocols are a research tool that are becoming very popular in language testing research, and so this first podcast for 20/11 focuses on its place in our Research Methods toolkit.

First, I asked Khaled to talk to me about his research, I then turned to Melissa Bowles, Assistant Professor of Spanish at the University of Illinois at Urbana-Champaign, to tell us more about the problems, pitfalls and possibilities of think-alouds in our research. Her recent book entitled *The Think-aloud Controversy in Second Language Research* was published in 2010 by Routledge.

Khaled, welcome to Language Testing bytes, and thank you for agreeing to talk to us about your article on think-aloud protocols.

Respondent: Thank you for inviting me to do this interview.

Interviewer: Khaled, first of all, can you very briefly tell us what think-aloud protocols are and what you have used them to research in your article?

Respondent: Think-alouds are basically a data collection procedure that asks the participants why they are doing a given task like writing, or rating, or reading, to say aloud what they were doing, thinking and feeling while performing the task, for example, writing or as I said reading to. The purpose of this study is to understand the processes and strategies that participants are using to do the particular task, such as, in this case, rating the essays.

This paper is actually based on a small part of a larger project that uses think-aloud protocols to examine… basically I wanted to compare novice and experienced raters [SL 02:15] of ESL writing, when they use holistic and then analytical writing scales, and to look at how that impacts or what differences are between the scores they assign but also looking at the differences in terms of their reading processes. And, for that, I use the think-aloud protocols with some participants, basically asked each participant to read a set of essays and think aloud, basically say what they were thinking or doing or feeling during the rating process from the point where they start reading the essay until they assign the final mark to it.

Interviewer: I think I first became aware of the kind of potential of think-aloud protocol analysis that you are now talking about when I read Gary Buck's 1991 paper in volume 8 of *Language Testing*, entitled The *Testing of Listening Comprehension*: *an introspective study*, where he asked to students to talk about how they responded to test items. It was clear from Buck's work that test takers arrived at answers to the same question in a variety of ways that couldn't really be predicted. Since then, think-alouds have been used in a variety of contexts many of which you summarise in your contribution to the Journal. Now, in your view, is your work and the work you describe trying to get to what Messick would have called substantive validity, by asking the question of what goes on people's heads when they take tests, or, in your case, when raters grade samples?

Respondent: Yes, it's interesting, and also the first time, actually, I learned about think-aloud protocols was Alister Cumming's article in 1990, about rater rating processes, but I mean it made me aware more of the usefulness of this technique. I mean, the question that the most [UI 04:09] using think-aloud protocols to examine raters behaviour tried to understand is, we know that raters assign scores, if they can assign the same score to the same essay for different reasons, raters are expected to refer to rating rubric or scale that specify what each score means, but most of the time there is also an element of judgment that means that basically the score assigned is based not on the description of the rating scale but on the interpretation the rater makes of that description, and also taking into account the writing task and the characteristics of the essay. So, yes, I mean, using think-aloud is one of the strategies that to understand the decision-making processes that raters go through.

Interviewer: And, in the light of your response there, what do you think are the main lessons that you've learned about the use of think-alouds as a research method in language testing?

Respondent: There are several things I think that I have learnt from this. The first is what I call differential impact of think-aloud protocols, meaning that maybe we know from the [UI 05:15] that think-aloud protocols has an impact on performance, but it's not clear from the literature what kind of impact this is, particularly with rater behaviour, although some research has been done on second language performance in general but not with rater behaviour and the impact of thinking aloud on that.

So, in this case it's basically saying that while the impact of think-aloud protocols maybe depends on a lot of factors depending on the characteristics of the individual, but also depending on the such contextual factors, in this case it was a rating scale. The second thing is the difference between reading aloud and thinking aloud. Most studies talk about thinking aloud but I found no reference to the impact of reading aloud versus reading silently, and it seems, which I report in this study, that reading aloud may have a different impact on different people in terms of the things that they pay attention to when reading the essays.

And the final point is, concerns I think the social dimension of think-aloud protocols, the way think-aloud protocols are referred to in the literature is that it's a monologue where the person is just talking, saying aloud what they are thinking. But, again, this study and also in other studies, although not in assessment, research shows that actually participants think about an audience when they are providing think-aloud protocols and that affects what they say and how they say it. And although this is thought of as a limitation I think it's true for all methods including, for example, interviews and experiments where there is that social dimension. I think the key is rather than looking at it as a limitation is actually to take it into account and then stand that that's part of the process or part of the context of doing think-aloud protocols.

Interviewer: Well, thank you very much for your insights based on your recent work, and thank you for joining us on Language Testing Bytes, Khaled.

Respondent: Thank you for inviting me again, it's been great talking to you. Thanks.

Interviewer: We're now joined by Melissa Bowles whose 20/10 book entitled *The Think Aloud Controversy*, has set out the problems, pitfalls and opportunities in using think-alouds in second language education and language testing. Welcome to Language Testing Bytes.

Respondent 2: Thank you for inviting me, I'm glad to be here.

Interviewer: To start, I understand that the two potential problems that researchers have to deal with in conducting think-aloud studies are verticality, if I'm pronouncing that

correctly, and reactivity, can you explain for us what these are and how they might impact on think-aloud research in the field of language testing?

Respondent 2: Certainly, as you said the two threats to the validity of think-alouds are verticality and reactivity, and I should start by saying that controversy has always surrounded verbal reports and that's both in cognitive psychology, where the method was first used, and in language research. And those two concerns are that verbal reports may, first, provide an inaccurate and/or incomplete account of thought processes, that's the issue of verticality, and, second, that they may have either a facilitative or detrimental effect on participants' processing and performance, and that's the issue of reactivity, and that's usually compared to participants who complete the same task silently, that's how that's determined.

In their classic book on verbal reports, Ericsson and Simon, regard verticality as a sort of unfortunate but unavoidable feature of verbal reports, and they comment that verbalisation can never truly capture all thoughts, and they say, in any case, that think-alouds can only reveal thoughts that are verbalisable, in other words those that are conscious, those that are encoded verbally, and how they work in memory. They also serve caution that think-alouds are more accurate and more complete than retrospective reports, which are done after the task is completed because of the lack of time delay in the think-alouds.

Reactivity, on the other hand, has been a source of considerable concern, and has been a topic of literally scores of empirical studies in cognitive psychology and more than a dozen now in second language acquisition. And findings have been all over the board, although the pattern that emerges indicates that thinking aloud slows task completion, so time and task, but does not tend to affect the outcome of task performance significantly so long as participants are asked only to think their thoughts aloud and not provide any sort of additional commentary or justify their decision during the task.

Interviewer: Now, turning to the article by Khaled Barkaoui, I was interested in the references to raters saying that, providing a think-aloud protocol often seem to interfere with the thinking process, I guess it seems to me that the moment you ask someone to think aloud you are going to put additional pressure on short-term memory and the ability to process information. Now, in your book you distinguish between non-cognitive and meta-cognitive think-alouds, what are these, and do they make a difference to performance?

Respondent 2: I should say that distinguishing between verbal reports by type is not my invention. Ericsson and Simon in their book on verbal reports also distinguish by level of reporting. They call the most basic type of report, Type 1, which is what I call non-meta-cognitive reports, which are those that require participants to verbalise only those thoughts that are going through their minds at the time that they're doing the task, that is, they require no additional commentary, no justification on the participants part. That's different from what Ericsson and Simon refer to as Type 2 and Type 3 reports, which I called meta-cognitive reports, and that reflects the idea that participants are justifying their thoughts commenting on their processing in that type of think-aloud.

And, certainly, where reactivity has been found it's most likely found with meta-cognitive reports, with those that require additional processing, additional justification, so that can certainly make a big difference.

Interviewer: And, do the kinds of tasks we ask raters or test takers to do make a difference to the quality of the data we get?

Respondent 2: Yes, it's clear that the task accompanying the think-aloud makes a difference. What's less clear is what the nature of that difference is so it's a bit of a problem. The reason for that is that most of the research that's been done on the validity of think-alouds, has been done in cognitive psychology with non-verbal problem solving tasks of various types, and Ericsson and Simon provide an excellent treatment of those studies in their 1993 book. My 2010 book provides a description and a meta-analysis of studies that have used think-alouds with verbal tasks, but again most of the research on the validity of think-alouds with verbal tasks were used during reading tasks.

So, the jury's still a bit out on exactly how type of task interacts with the think-aloud because we simply don't have enough data points on think-alouds with different types of tasks. But, in general, we can say, I think, both based on the cognitive psychology research findings and on my findings, think-aloud protocols are more likely to be reactive when they're used with more complex tasks, that are more demanding, or involve multiple factors to solve or to accomplish, than when they're used with less complex tasks or those that involve either one or a small number of factors to solve or to accomplish. So, a lot more research needs to be done in that area.

Interviewer: Before we come to the end of our time on this podcast, let's move on to findings. In the study by Barkaoui I was interested to read the conclusion that, and I quote here, "Data indicated that the thinking aloud affected several participants rating processes, rating criteria, and/or the scores they assigned. These effects seemed also to vary across raters and rating scales", and that's the end of the quote. This may suggest, could suggest, that the effects of this research tool are fairly unpredictable, which doesn't really sound terribly optimistic. From your research can you summarise what we know about the effects of think-alouds on participants, and perhaps speculate just how useful think-alouds might be in future language testing research?

Respondent 2: Okay, the variables that interact to cause reactivity are right now not very well understood but they're thought to include things such as the time or reporting, which I mentioned, so whether the report is concurrent to the task or, of course, after the task; the type of report that meta-cognitive versus non-meta-cognitive distinction; the type of task; the language of the task; the language of verbilisation, so the language that the participant doing the think-aloud is speaking, whether that's their first language, their second language, some other language or a combination of those.

In my book, which reports on a quantitative meta-analysis of reactivity in studies that involved verbal tasks, I found that think-aloud groups did not consistently perform significantly differently than the silent control groups. Results on time were more decisive and those indicated, pretty much across the board, that thinking aloud increased time on task compared to silent task completion. From that, I would conclude that inferences made on the basis on verbal reports should always be taken cautiously and that, whenever possible, a small, silent control group should be included in studies that use the think-aloud measure. So, that that way at a minimum the verbalisation group's performance can be compared to that of an otherwise matched group that completes the same task silently, and that way it's a sort of check on reactivity on a study by study basis.

By the same token I think if researchers follow guidelines for administration of think-aloud protocols they can be a valid data collection method that gives us insight on what otherwise we wouldn't be able to know. In that way I see think-alouds as a window into cognitive processes, and in that way they provide a glimpse

into the mind, even though we know it doesn't provide a complete look, it's just a glimpse into the mind, but at least it gives us a little bit of knowledge that otherwise we wouldn't have.

In language testing I think, certainly Barkaoui's paper shows how think-alouds can provide insights on expert and novice rater behaviour, and previous studies in language testing have also used think-alouds to look at examinees taking tests as a way of gauging their strategy use. So, I see a lot of potential in language testing for introspective measures like think-alouds, I think we just, we need to be a little bit careful in the way that we implement them and interpret their findings, but they can provide us with some data that otherwise we would have no way of getting at.

Interviewer: Well, Melissa, thank you very much for joining us and joining us on Language Testing Bytes, and for sharing your expertise on this fascinating area of research with us.

Respondent 2: Thank you very much for giving me this opportunity to speak to the audience of language testing.

Interviewer: Thank you for listening to this issue of Language Testing Bytes.

Language Testing Bytes is a production of the *Journal of Language Testing* from Sage Publications.

You can subscribe to Language Testing Bytes through iTunes, or you can download future issues from ltj.sagepub.com, or from languagetesting.info.

So, until next time we hope you enjoy the current issue of *Language Testing*.

[End of recorded material]

NOTES: [SL 00:00] Sounds Like

[UI 00:00] Unintelligible