THE EDGE OF EXPERTISE: TEST WRITERS THINK ALOUD

Katy Salisbury

University of St Mark and St John (UNITED KINGDOM)

Abstract

What does it mean to be an expert within a given field? This paper gives an overview of a small-scale research study which explores this question through the prism of one very specific domain – listening test item writing in TESOL (Teaching English to Speakers of Other Languages). Considerable attention has been given by different researchers to *outcomes* - the ESOL tests themselves - but very little consideration has been given to *process*: how the tests are produced by item writers. This research uses frameworks frequently found in expertise studies (see, for example, Ericsson et al. 2006) and concurrent-verbalisation (think-aloud) data-gathering techniques (for example, Bowles, 2010) to illuminate processes involved in listening test item writing. It compares the way small groups of novice and experienced test writers undertake the same item-writing task. The findings indicate that, although expertise in test writing is highly individualised, a number of identifiable strategies and practices are associated with expert outcomes.

The paper will summarize the ways in which the insights from this study have the potential to inform ESOL test-writer training and it will also make an introductory case for using concurrent verbalization to shed light on the cognitive operations used by experts in many different domains.

Keywords: ESOL test writing, test-writer training, think-aloud; verbal protocol analysis (VPA).

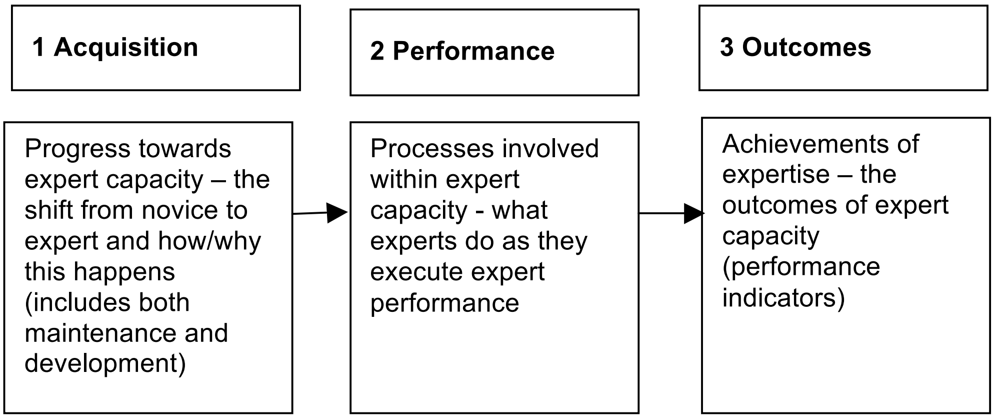
1 INTRODUCTION

What do expert test writers do which non-experts do not? What gives experts the edge? Why do some people take to test writing, whilst others find it deeply uncongenial? I had been a teacher for some years when two colleagues and I were ask to try our hand at ESOL test creation. My colleagues were unimpressed, saying the exercise was reductive and frustrating. My own response was strikingly different: I found the process fascinating and felt it engaged hitherto untapped skills. Over the past 25 years I have continued test writing (also known as item writing) on a freelance basis alongside my main teaching job, regularly undertaking commissions for different testing bodies. I have become interested in the possible reasons for these radically different reactions to the process of test writing and I wanted to find out more about the cognitive operations used by people who find success in the domain, focusing on expertise in listening test production. Item-writing for large examining bodies tends to be organized on a 'cottage industry' basis, where the job is outsourced to freelancers working in their own homes. Although writers meet at certain junctures in the test-production cycle, many key stages are undertaken in isolation. I felt that it would be enlightening to open up some of the practices within this hitherto-closed professional sub-community.

Expertise is a much researched issue: see for example the collection of reports in Ericsson et al. (2006), which analyze what experts do in a wide range of domains, from chess playing to medical diagnosis. However, there has been comparatively little such research in TESOL, with just one dedicated collection of articles, published in 2005: '*Expertise in Second Language Learning and teaching*'. In his introduction to this book, Johnson says the time is ripe for growth in TESOL expertise studies suggesting they will bring 'huge possible benefit to language learner and teacher training' (2005:1). In this study I seek to contribute in a small way to this body of work, using conceptualizations and processes widely used in expertise studies to shed light on test-writing processes.

2 BACKGROUND

In any domain, there are three major elements which might be said to constitute a specification of expertise (see Fig.1). Element 1 is *Acquisition:* how practitioners move from novice to expert in the domain. Element 2 is *Performance:* the processes experts use while they undertake tasks. Element 3 is *Outcomes:* what experts actually achieve having performed the tasks. (See, for example, Anderson, 2009). In my target domain, much attention has been paid to Outcomes: the extent to which listening tests are valid, reliable and practicable (for example, Geranpayeh & Taylor, 2013), but my focus in the present study is Element 2 ***Performance*** processes: what expert writers actually do as they go about devising listening tests. I am also very interested in Element 1 - how expert item writers acquire their capacity - but that has been the subject of a separate study (see Salisbury, 2005).



1 Acquisition

Progress towards expert capacity – the shift from novice to expert and how/why this happens (includes both maintenance and development)

2 Performance

Processes involved within expert capacity - what experts do as they execute expert performance

3 Outcomes

Achievements of expertise – the outcomes of expert capacity (performance indicators)

Fig.1 Three elements of expertise

In a seminal paper written in 1988, Glaser and Chi identified a number of generic features of how expert practitioners in different fields perform their tasks (or solve problems), which I summarize as follows. Within their own domains, experts have what might be called ‘knowledge-organizing knowledge’. They select what information to memorize; store and retrieve it effectively (often in chunks - stored units formed from integrating smaller pieces of information) and use it to decompose a problem into manageable sub-problems; and then proceed to solve the problem, producing the desired outcome. When initially presented with a problem, these experts spend time trying to understand its nature, which can slow down their performance in the earlier stages but they do this in the knowledge that it will eventually lead to superior outcomes. Although experts’ memories are no larger than novices’, they use their memories more effectively within the domain because automaticity of portions of domain-specific skills frees up resources for greater storage.

To this list of generic features of expertise I the issue of avoidance of inappropriate difference reduction (Anderson, 2009). This difference reduction is often also called ‘hill-climbing’ because of the frequently-cited example of walkers trying to reach the top of a mountain. They do not have a map and go up the first incline they encounter but discover that they are climbing a hill lower than their target one: what seems to be difference *reduction* can often lead to difference *increase*, requiring painful additional effort to reach their objective. Experts recognize early on when this is happening and identify less direct but ultimately more suitable routes to their goal.

Of more specific relevance to my target domain is research done by Johnson (2003) on language teaching task design. The first major area which distinguishes expert designers from novices is what he calls their *logistical control.* Experts analyze their brief (the problem specification) with meticulous care and identify early on what is salient. They use a breadth-first rather than a depth-first approach, which means they survey several possibilities before plunging into working out details. As they design, they look ahead: envisaging possibilities and simulating what users of the teaching tasks might say and think. They proceed opportunistically, dealing with other sub-problems as they arise incidentally to the main work in hand. The second major area in which expert task designers are distinguished from novices is what Johnson calls *enrichment*: how they ensure that their task is ‘sufficiently detailed and rich’ (2003:128). He found that experts show highly developed awareness of the mechanics of implementing the tasks in the classroom, giving attention to a wide range of variables so their tasks can accommodate a wide range of potential contexts.

Johnson, like many researchers in expertise studies, used concurrent verbalization to gather data for his studies. This approach is also known as ‘think aloud’, and draws on the work on verbal protocol analysis (VPA) by Newell and Simon (1972). As research participants undertake a particular task, they are required to say out loud what is going through their heads. This talk is usually captured in audio recordings and transcribed to produce verbal protocols (VPs), which are then subjected to analysis. A key concept in cognitive science is *veridicality* (Bowles 2010): the degree to which perception accurately represents ‘reality’. For many researchers elicitation techniques such as interviews are not felt to reflect accurately on-task thought processes because participants ‘simply may not recall what they were thinking as they completed the given task' (Ibid.:14). Think-aloud is believed to enhance veridicality by associating task performance and report (verbalization) more closely.

Another feature of many expertise research designs is the comparison between experienced practitioners and novices. The word 'expert' is cognate with the word 'experience' (from the Latin *expertus* - past participle of *experiri*, meaning *try).* When placed, as it often is, in opposition to 'novice' (Latin *novicius* from *novus* – *new*), the equation of expert with experience is reinforced. Much has been written about the Ten Year Rule (for example, Ericsson et al, 2006:327): the notion of ten years and 10,000 hours' experience as a requirement for mastery in any domain. However, meta-research by Ericsson indicates that longevity is not in itself sufficient: 'There is now ample evidence from many different domains that the number of years of experience is a poor predictor of objective professional performance' (2009:2). This has led many writers, notably Bereiter and Scardamalia (1993, inter alia), to highlight the importance of distinguishing 'true' experts from experienced non-experts.

3 RESEARCH DESIGN

With these frameworks in mind I specified three questions which guided my research:

1 What micro-processes do expert listening test item writers use?

2 How does *expert* test-writing performance differ from *non-expert* performance?

3 How might insights gained from this research impact on listening test writing practice?

I identified ten research participants and invited them to undertake a particular test-writing task. Five had substantial experience of listening test item writing. The other five had no experience of test writing but did have knowledge of the domain of language assessment, through their own learning and teaching experience. I did not assume that the experienced writers would necessarily out-perform the novices and asked experienced test editors to rate 'outcomes' i.e. the quality of the tests the participants devised. The editors' ratings broadly endorsed the 'experience => expertise' equation, though it is important to note there were two significant 'outliers': one novice who scored highly and one experienced writer who received a relatively poor rating. On the basis of these editors' ratings I designated two participant groups: five 'experts' and five 'non-experts'.

Both groups were set the same test-writing task, which closely resembled what is normally required in this domain. The main components of this task are summarized as follows:

- Read a given newspaper feature article
- Review the test-writing specifications – what is required by the testing body
- Use these specifications to devise a listening proficiency test for CEFR C1 level:
  - Create a script by adapting the written text (with a view to this being read aloud by an actor)
  - Write 8 'objective' items (multiple-choice questions or 1-3-word gap-fill)
- Take up to 90 minutes to do this task

Whilst doing this task, participants were required to think-aloud, with the following instructions: 'As you work on the test, say aloud into the audio recorder everything that crosses your mind. Try not to worry how it might sound to me. Free associate as much as you wish and don't worry if what you say doesn't sound logical.'

4 CODING

Having audio recorded and then transcribed what the participants said (their VPs), I used a modified grounded-theory approach for the analysis, attempting to build theory directly out of data, rather than from preconceived concepts or hypotheses (Charmaz, 2011). As I transcribed and then read and re-read the VPs I began to identify patterns and to generate codes for sections of VP. Through a recursive process between macro- and micro-coding, I began to 'render' the data, writing memos on configurations emerging and moving towards more generative levels of granularity and applicability (Corbin and Strauss, 2008). I view the application of grounded theory as a taxonomical exercise, similar to the Linnaean system of arranging and labelling *genera*, *species* and *phyla* of the animal kingdom; it is the very act of naming that *builds* theory, and it the very act of 'grounding' that converts them into *communicable* theory.

At a macro-level, I identified five major stages in the VPs, which all participants went through and which I called episodes:

- Review specs (i.e. review test specifications)
- Read text
- Devise context
- Devise script
- Devise items

At a micro-level, I identified separate mental operations or 'cognitive operators'. I then applied codes according to what I understood the participants did, to what object, with what reason; I designated these three as 'VERB, object and *justifier' (*with different type-faces to distinguish them). For example, in the following VP extract in Table 1, (from her Review specs episode), an expert participant is making a comment on the test specifications as she reads them, noting that it is important to consider how accessible the text is for exam candidates in cultural terms. This was coded as follows. (A full list of my Operator codes is given in the Appendix at the end of the paper.)

Table 1 Sample operator coding

VP (think-aloud transcription) (segmented/)	**Episode**	**Operator code**
/'contents should be accessible to candidates of different ages and nationalities' [reads specifications] /something I'll have to bear in mind thinking of that…don't want it to be too ethnocentric so think carefully about references to …um…features which may be beyond the scope of some candidates for the examination……./	Review specs	READ specs COMMENT specs *access:culture*

5 ANALYSIS

From an analysis of all the VPs, I identified two striking **macro** patterns in the data. Firstly, for experts the Devise items episode comes before their Devise script episode: while the order is reversed for non-experts. Secondly, experts embed episodes much more than non-experts, e.g. beginning to devise items while they are in the process of reviewing specifications.

My analysis at **micro** level also showed clear distinguishing patterns. The following quantitative analysis (Table 2) gives more-detailed information about the operators contained within each episode (figures are given as percentages of total operators of each participant). The table shows that for all participants there is a much greater emphasis on Devise script and Devise items than on the other episodes. Generally, however, experts show more equal distribution of operators than non-experts, with no *one* episode strongly favoured at the expense of others. Notably, the Devise context and Review text episodes are given significantly more attention by experts than by non-experts.

Table 2 Percentages of operators occurring in each episode

	Participant (code name)	Review text operators %	Review specs operators %	Devise context operators %	Devise script operators %	Devise items operators %
Non-expert	**Emily**	3	8	0	24	65
	Gary	2	2	1	21	74
	Rory	13	11	3	40	33
	Teresa	6	2	3	62	27
	Zach	13	3	9	58	17
Expert	**Anne**	6	4	3	69	18
	Caitlin	22	12	11	21	34
	Joe	12	2	6	39	41
	Sharon	2	5	8	42	43
	Malcolm	17	8	5	45	25

As an example of my qualitative analysis of **micro-coding** I compare the operators used by a non-expert (Emily – Table 3) and an expert (Caitlin – Table 4) in their respective Review specs episodes.

Table 3 Non-expert Emily Review specs operator codes

Segments	Operators
1.	COMMENT specs:text type COMMENT specs:text length COMMENT specs:access COMMENT specs:marking -ve EML *financial constraints*
2.	CONTROL:PROCEDURE rubric *writer facility (*NF)
3.	CONSIDER item type COMMENT –ve RTL item type *sub-skill focus (*X) COMMENT –ve self item type PROPOSE item type (NF)

Non-expert Emily's micro-processing appears relatively rich and complex, with three segments and a total of nine operators, covering a variety of aspects (five distinct facets). However, the majority are lower-order COMMENT specs, and although she includes three justifiers, one is erroneously cited (X) and there are two proposals which although posited relatively firmly here, are not followed through (NF).

By contrast, the expert Caitlin's Review specs episode (Table 4) comprises five segments and considerably more operators (21).

Table 4 Expert Caitlin Review specs operator codes

1.	READ specs COMMENT specs:text length COMMENT specs:text type CONTROL note and recursion
2.	READ specs READ text CONSIDER context:text type *repertoire TP* CONSIDER context:text type *access culture* CONTROL recursion
3.	COMMENT specs:access culture CONTROL procedure COMMENT specs:language *level/access* READ specs COMMENT specs:*access culture*
4.	READ specs:marking non-expert COMMENT specs:marking correct spelling COMMENT specs:authenticity MODIFY context setting (recursion) *authenticity:contextual* COMMENT specs:permissible change
5.	READ specs:required products CONSIDER context: setting *authenticity:contextual*

As Caitlin starts to review the specifications she immediately decides on a strategy for recording her observations and ideas (CONTROL note). As she reads the specifications she comments on seven facets. Her COMMENTs are interspersed with CONSIDER operators, notably of context:text type and setting. This goes beyond simple description (that is, she tends to avoid the simple READ or COMMENT operators). Caitlin uses this episode to explore higher-order issues as she reads the specifications. There is significant multi-tasking as she reads and considers options, shown by the variety of different operator codes.

6 DISCUSSION

From this combination of quantitative and qualitative and macro and micro analysis it was clear that there was considerable variation in the way expert participants went about the task, leading to the conclusion that performance processing in listening test item writing is highly individualized. However, a number of core characteristics are identified as representative of expert (as compared to non-expert) performance in this domain, as follows.

Expert test writers have more effective *internalization of task (problem) elements*, needing fewer checks on specifications and making fewer mistakes in following them. They *multi-task* through episode embedding, for example, they use several different episodes throughout the whole VP to enrich their understanding of the task. They are aided in this by an *effective working memory*, retrieving decisions across episodes. This reflects both Glaser and Chi's (1988) and Johnson's (2003) findings that experts conceptualize problems in a semantic rather than syntactic way and have greater facility in using memorized material.

A second area of distinctive performance relates to the time and effort experts spend on the different task elements. In particular, as mentioned above, they place significantly greater emphasis on *context*, taking a longer time to instantiate it (both within the Devise context episode and incidentally as the need arises in other episodes). They also encompass many different aspects of context - setting, topic, field, tenor and mode - while non-experts tend to confine themselves just to setting and topic. This could be said to be, as was the case with Johnson's task designers, a breadth-first approach, taking time to set the scene before engaging at a detailed level with content. It also shows a greater awareness of the end-users' needs i.e. giving necessary background to help listeners understand the text as discourse. Another finding is that experts spread their *attention in a more balanced way than non-experts* and avoid obsessive concern with relatively superficial issues such as text length or item type. They show a greater understanding of the reasons for taking actions with a greater density and variety of verbs, referents and justifiers.

Finally, the fact that experts devise *items before text* is very telling. I see this as an example of avoidance of inappropriate difference reduction. Non-experts appear to follow communicative teaching and materials design precepts and privilege the script. They tend to retain the original wording of the text wherever possible but nearly always have to modify this after they have written the items. By contrast, experts recognize that in tests, particularly of listening, items need to be carefully spaced – by devising items first they remove the need for a later script rewriting stage. Experts have a flexible, almost iconoclastic, confidence which privileges items over text in the service of efficient test design.

A number of these expert characteristics may be directly attributable to greater knowledge of the domain and to repeated practice within it. However, the fact that these behaviours are exhibited by the one inexperienced expert but less so by the one experienced non-expert suggest these actions are not simply a function of longevity in working in the domain.

7 REFLECTIONS

An expertise study using verbal protocol analyses makes considerable demands on the researcher. It takes a great deal of time to transcribe and analyze such long recordings of detailed musings in 'think-aloud' mode. However, I feel the effort is worthwhile on a number of different levels.

Firstly, the method appears to be an effective means of capturing cognitive micro-processes. It is beyond the remit of this small-scale study to determine veridicality but my participants reported that they were comfortable 'thinking aloud', that they quickly forgot they were doing it and that it did not significantly affect their task performance. This would seem to counter one key argument against using think-aloud – its potential *reactivity*: 'acting as an additional task and altering cognitive processes rather than providing a true reflection of thoughts' (Bowles, 2010:14).

Secondly, the actual process of close, line-by-line coding is of significant value. Although painstaking and time consuming, doing it gives the researchers an intimacy with the data which is not possible through more macro-level analysis and prevents them from projecting too forcibly their own concerns onto the data.

Thirdly, although the study was very small-scale and exploratory in nature, a number of interim findings (summarized above) may be of value for the broader community of listening test item writers, particularly when designing training sessions, giving a baseline from which to consider the most effective induction of new writers and to support the development of existing writers. This approach does have its detractors, who claim it is anti-developmental to use existing practices as a model to be replicated by novices because it follows the 'discredited' Craft Model of professional development (see Wallace, 1991). However, I feel these findings should be viewed as a starting point for individual learning. I also feel the codes which emerged from the study (see Appendix) represent a valuable outcome in themselves. They constitute a common 'vocabulary' of domain terms and a common 'syntax' of relations between them. In other words, they provide a consistent language for defining, describing and exemplifying operations within the domain of listening test item writing and provide a resource from which to sample, and a baseline from which to critique practice.

I accept that this study has a number of limitations because of its small scale and highly controlled experimental design: the 90-minute time restriction and the specified text and task. As Bereiter and Scardamalia (1993) point out, many of the most interesting insights about expert behaviour reveal themselves over a much longer time period, for example, how experts 'incubate' an intractable problem or use learning from a given problem to inform their solution to a different one. Bereiter and Scardamalia also talk about the phenomenon of experts seeking opportunities to work at the 'edge' of their competence, constantly looking at ways to 'complexify' problems as a means of enhancing, or even of maintaining, their skills and understanding. This is often revealed in experts' capacity for seeing 'promisingness' in unusual base texts or in their lateralizing capacity about the choice of test focus and item type. For further discussion of these elements, see Salisbury's (2005) naturalistic study of expert test writers, which forms a companion piece to the present experimental research.

To conclude, I feel this small-scale project endorses the value of expertise research, showing how it can enable members of a professional community to bring about positive change through increased awareness of what they do and why they do it: in short, to theorize from practice. I believe it has the potential to bring to the surface the heuristics being used by expert practitioners and make them available for scrutiny and as a basis for learning, both within the domain and beyond.

REFERENCES

[1] Ericsson, K.A., N. Charness, P.J. Feltovich and R.R. Hoffman (eds.). 2006. The Cambridge Handbook of Expertise and Expert Performance. Cambridge: Cambridge University Press

[2] Bowles, M. 2010. The Think-Aloud Controversy in Second Language Research. New York: Routledge

[3] Johnson, K. (ed.) 2005. Expertise in Second Language Learning and Teaching. Basingstoke: Macmillan.

[4] Geranpayeh, A. and L. Taylor. (eds.). 2013. Examining Listening: Research and Practice in assessing second language listening. Cambridge: Cambridge University Press

[5] Salisbury, K. 2005. The Edge of Expertise? Towards an understanding of listening test item writing as professional practice. Unpublished PhD thesis. King's College University of London

[6] Glaser, R. and M.T. Chi. 1988. Overview. In Chi, M.T., Glaser, R. and M.J. Farr. (eds.). 1988. The Nature of Expertise. Hillsdale, NJ: Lawrence Erlbaum Associates.

[7] Anderson J.R. 2009. Cognitive Psychology and its Implications. 7th Edition. New York: Worth Publishers

[8] Johnson, K. 2003. Designing Language Teaching Tasks. Basingstoke: Macmillan.

[9] Newell, A. and H.A. Simon. 1972. Human Problem Solving. Englewood Cliffs, NJ: Prentice Hall. Ericsson, K.A. (ed.). 2009. Development of Professional Expertise. Cambridge: Cambridge University Press

[10] Charmaz, K. 2011. Grounded Theory Methods in Social Justice Research. In N.K. Denzin and Y.S. Lincoln (eds) *The Sage Handbook of Qualitative Research* (4th Edition). Thousand Oaks, Calif: Sage

[11] Corbin, J. and A.L. Strauss 2008. Basics of Qualitative Research: Grounded Theory Procedures and Technique (3rd Edition). Newbury Park, CA: Sage Publications Ltd.

[12] Bereiter, C. and M. Scardamalia. 1993. Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise. Chicago: Open Court Publications.

[13] Wallace, M.J. 1999. Training Foreign Language Teachers. Cambridge: Cambridge University Press.

APPENDIX OPERATOR CODE LIST

	Verbs		Objects
1	COMMENT	1	context
2	CONFIRM	2	editing procedure
3	CONSIDER	3	IBT/ IBC
4	CONTROL	4	item distribution
5	DEFER	5	item focus
6	DISTIL	6	item type
7	METACOMMENT	7	item wording
8	MODIFY	8	key
9	PARAPHRASE	9	key wording
10	PROPOSE	10	language
11	READ /RE-READ (LISTEN)	11	main text point/gist
12	REHEARSE/ROLE PLAY	12	rubric
13	REJECT	13	rubric wording
14	REVIEW	14	script
15	TRANSFUSE	15	specifications
		16	structure
		17	sub-skill area
		18	text
		19	text content (HO/LO)
		20	text cut
		21	text type

	Justifiers		
1	*access*	*18*	*intrinsic interest*
2	*authenticity*	*19*	*key constraints*
3	*candidate empathy*	*20*	*key givens*
4	*balance*	*21*	*level*
5	*coherence*	*22*	*orientation*
6	*cohesion*	*23*	*permissible item type*
7	*core concept*	*24*	*personal ethics*
8	*discrimination*	*25*	*personal resonance*
9	*distinguishability*	*26*	*process requirements*
10	*distraction*	*27*	*repertoire (teaching/TP/TIW/Lg*
11	*editing considerations*	*28*	*rhetorical function*
12	*exploitability*	*29*	*sensitivity*
13	*factual correctness*	*30*	*shelf life*
14	*financial constraints*	*31*	*spellability*
15	*gradation*	*32*	*sub-skill focus*
16	*guessability*		
17	*international dimension*		