Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles

Luke Harding, J. Charles Alderson and Tineke Brunfaut

Lancaster University, UK

Abstract

Alderson, Brunfaut and Harding (2014) recently investigated how diagnosis is practised across a range of professions in order to develop a tentative framework for a theory of diagnosis in second or foreign language (SFL) assessment. In articulating this framework, a set of five broad principles were proposed, encompassing the entire enterprise of diagnostic assessment. However, there remain questions about how best to implement these principles in practice, particularly in identifying learners' strengths and weaknesses in the less well-documented areas of SFL reading and listening. In this paper, we elaborate on the set of principles by first outlining the stages of a diagnostic process built on these principles, and then discussing the implications of this process for the diagnostic assessment of reading and listening. In doing so, we will not only elaborate on the theory of diagnosis with respect to its application in the assessment of these skills, but also discuss the ways in which each construct might be defined and operationalized for diagnostic purposes.

Alderson, Brunfaut and Harding (2014)[1] have recently applied the lessons learnt from studying how diagnosis is practised in other professions to outline a theory of diagnosis

Corresponding author:

Luke Harding, Department of Linguistics and English Language, Lancaster University, Bailrigg, Lancaster, LA1 4YL, UK.
Email: l.harding@lancaster.ac.uk

in second and foreign language (SFL) assessment. A theory of this kind would necessarily contain: a robust definition of diagnosis, a clear understanding of the means of diagnosis and the participants involved, recommendations for a set of procedures for conducting diagnoses and a closer focus on the interface between a decision, the development of appropriate related feedback and the intervention to follow. In mapping out this theory, Alderson et al. (2014) developed a tentative framework which can be translated into a set of five principles for the diagnosis of strengths and weaknesses in SFL (pp. 21–22):

1. It is not the test which diagnoses, it is the user of the test.
2. Instruments themselves should be designed to be user-friendly, targeted, discrete and efficient in order to assist the teacher in making a diagnosis. Diagnostic tests should be suitable for administration in the classroom, designed or assembled (with recourse to existing suites of tools) by a trained classroom teacher (or other experienced language teaching professional), and should generate rich and detailed feedback for the test-taker. Most importantly, useful testing instruments need to be designed with a specific diagnostic purpose in mind. This principle is derived from the emphasis the interviewees placed on tools with a clear focus and capacity to play a facilitating role.
3. The diagnostic assessment process should include diverse stakeholder views, including learners' self-assessments.
4. Diagnostic assessment should ideally be embedded within a system that allows for all four diagnostic stages: (1) listening/observation, (2) initial assessment, (3) use of tools, tests, expert help, and (4) decision-making. Much current diagnostic testing arguably begins at stage (3), using general diagnostic tests for whole populations rather than more targeted measures that have been selected on the basis of stages (1) and (2). … A theory of diagnosis should not preclude large-scale assessments, but it should also pose a challenge to these programmes: would the same decisions about strengths and weaknesses have been made on the basis of an individualized assessment in a classroom context?
5. Diagnostic assessment should relate, if at all possible, to some future treatment.

One of the key practical outcomes of this set of principles is the proposal for an "ideal" diagnostic process in principle 4 which is illustrated and expanded on in Figure 1. Note that this four-stage process addresses principles 1, 2, 3 and 5 in its recommendations for: stake-holder involvement (including learners themselves) *(principle 3)*; targeted, purpose-built diagnostic tools, selected from a bank according to purpose *(principle 2)*; rich and detailed feedback *(principle 2)*; and treatment or intervention to address specific problems which have been identified *(principle 5)*. All of this is performed by a skilled "diagnostician"[2] *(principle 1)*.

While a process of this kind has a firm theoretical basis, having been abstracted from descriptions of diagnostic practices across a range of professions (including such fields as medicine, education, information technology and mechanics; see Alderson et al., 2014), its application to the field of language assessment remains untested. This is not
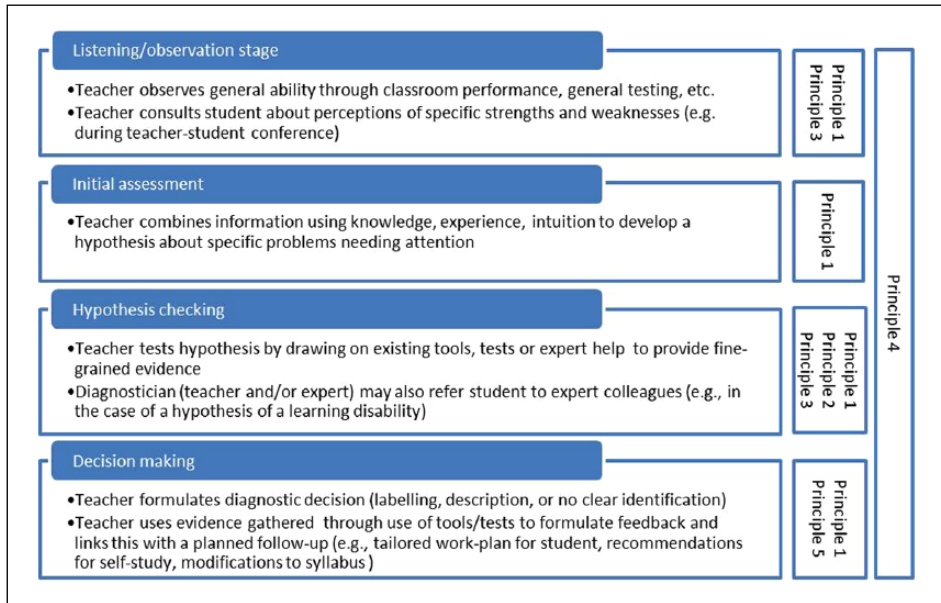
Figure 1. The diagnostic process.

to say, however, that these principles might not be easily applicable to some existing assessment practices, diagnostic or otherwise. For example, many of the routine procedures involved in second language writing assessment, for instance those in portfolio assessment contexts, might be seen to fit with this set of procedures.

Much less has been written, however, about the ways in which a process of this kind might be implemented in order to provide diagnostic information about the receptive skills: reading and listening. This may be for several reasons. Reading has received some attention, but much more in the areas of L1 reading than SFL reading (Grabe, 2009). Listening in particular is an under-researched skill in general, let alone in the contexts of classroom-assessment (Vandergrift & Goh, 2012). Studies of listening pedagogy in the classroom have noted that listening instruction tends to follow a "testing model", where listening tasks are performed, answers are checked, and little help is provided to students in terms of how to improve their listening (Field, 1998, 2008). It can be assumed that much classroom listening assessment is therefore akin to large-scale proficiency testing conducted in a classroom environment (but see the pedagogical model proposed by Vandergrift & Goh, 2012).

The aim of this paper, then, is to explore the implications of our proposed diagnostic process for the assessment of reading and listening. In particular, we will concentrate on the types of tools and tests that might usefully be developed to be used within the system described in Figure 1. In doing so, we will not only elaborate on the theory of diagnosis with respect to its application in the assessment of these skills, but also discuss the ways in which each construct might be defined and operationalized for diagnostic purposes.

The study will discuss reading first, and then listening, before concluding with some general statements about where further research might best be directed. Both the section on diagnosing reading and that on diagnosing listening will be guided by the nature of these macro-skills. The section on diagnostic assessment in SFL reading will introduce a range of broader issues regarding diagnostic assessment, while in the listening section, a worked example of how a particular listening problem might be further investigated within the diagnostic process is presented. Unfortunately, due to space limitations it was not possible to describe a similar hypothetical example for SFL reading, but many aspects of the listening example directly apply to diagnosing SFL reading. Similarly, core issues raised in the section on SFL reading are also relevant to SFL listening.

Diagnostic assessment in SFL reading

The SFL reading construct

In the diagnosis of first language reading problems, there is a huge literature of theories and research to draw upon (Carver, 2000; Clay, 1985; Ehri & Snowling, 2004; Guthrie & Wigfield, 1997; Kintsch, 1998; Perfetti, 1985; Snowling & Hulme, 2005). Not so in SFL reading, although the number of publications, studies and theories has grown in the last two decades (e.g., Alderson, 2000; Grabe, 2009; Koda, 2005). Much of the SFL literature is dependent on or derivative of the L1 literature. Whilst there are doubtless many common areas across these two fields, there are, of course, also important differences, and we would argue that it is more important to understand the differences, since the commonalities have been well researched in first language reading studies and have been carried over to the SFL reading literature. That is, however, not to say that it is not important to explore whether the sorts of diagnostic measures that are common in L1 reading can be useful in SFL reading also, a topic which has been explored in the DIALUKI project (www.jyu.fi/dialuki) and Alderson, Haapakangas, Huhta, Nieminen and Ullakonoja (2015).

The obvious and probably most important differences between the two fields lie in language. SFL readers are typically reading in a language that they have not mastered, and therefore L2 reading problems will be at least as much language-related as reading-related (Alderson, 1984). Therefore diagnosing SFL reading will involve diagnosing language problems as well as strictly reading problems. In the case of L1 reading, diagnosis usually assumes that the reader already knows the spoken version of the language and "simply" has to learn how to understand the written version. This is not to say that the written and spoken languages are the same, which they are clearly not, but it does mean that L1 readers initially have to learn how to recognize in the written form what they already know in the spoken form.

Nevertheless, it is probably reasonable to accept that both L1 and SFL reading involve a number of different "levels" of ability. The lower level is that of word recognition, of being able to link written symbols to sounds or meanings (depending on the orthography of the target language), the parsing of syntactic and morphological structures, the development of the automaticity of such recognition and parsing, and the role of working memory and attention in deriving meaning from text. The higher level is more diffuse, but involves skills and resources such as topic and world knowledge, inferencing,

building a mental model of the text, monitoring what one has read and what or whether one has 'understood' it and synthesizing and evaluating what one is reading and has read (Grabe, 2009).

A distinction is often made between a text model of what the author has written and a situation model of what the reader has understood, that is, an interpretation of the text in light of one's topic and world knowledge and one's purpose in reading. Purpose in reading also affects the type of reading one does, what has been referred to by some as 'expeditious' reading, such as skimming, scanning and search reading (Khalifa & Weir, 2009) or careful reading at both the local and global levels. This distinction is also used by Carver (1985, 1992) who describes the different speeds at which readers read for different purposes.

Diagnosing lower-level SFL reading processes

Drawing on the L1 literature, it is easier to see how one might be able to diagnose SFL reading at the lower level than at the higher level. The cognitive processes involved in word recognition are fairly well understood and numerous measures have been developed to test such processes. The DIALUKI project mentioned above, for example, used backward digit span, rapid word reading, rapid word list reading, rapid automatized naming, non-word reading, non-word spelling, non-word repetition, phoneme deletion and common unit tests. An interesting innovation was to use these measures of cognitive processes not only in L1 Finnish (a transparent orthography) but also in FL English (which has an opaque orthography). It was discovered that the FL versions of the cognitive measures accounted for much more of the variance in FL English reading than did the L1 versions of the same measures. It was therefore concluded, in contrast to the usual advice to administer such cognitive measures in L1, that in the SFL context it was more useful to use such cognitive measures in the SFL.

Indeed, going beyond the word recognition and phonological decoding measures, we venture to suggest that the diagnostic testing of SFL reading should always use FL measures, but take into account the L1 of the learner. Thus, for example, measures of syntactic and morphological parsing, also held to be lower-level abilities that need to be automatized for fluency in SFL reading, should take account of the differences between the learner's L1 syntax and morphology and that of the target language. This has the consequence that global diagnostic measures that do not take account of the learners' L1 are likely to be much less useful than are those which take such differences into account.

Diagnosing higher-level SFL reading processes

The high-level processes and resources are, of course, also important in SFL reading, but much less is known in detail about how to diagnose weaknesses in these areas. For example, it is common practice to identify different so-called components of reading, and, to the extent possible, to test separately the ability to understand gist, the ability to understand the main idea, to understand specific detail and so on. For example, DIALANG (www.lancaster.ac.uk/researchenterprise/dialang/about.htm) tests the three reading subskills "understanding the main idea", "making inferences" and "understanding specific

	Subskills tested	Text type	Theme
✗	Identifying specific information	News article	Animal ethics
✗	Understanding main ideas and supporting ideas	Academic article	History and culture
✗	Understanding information and making an inference	News article	Environmental issue
✗	Interpreting an attitude or intention of the writer	Fiction	History and culture
✗	Understanding main ideas and supporting ideas	Academic article	History and culture
✗	Interpreting a word or phrase as used by the writer	Academic article	History and culture
✓	Identifying specific information	News article	Environmental issue
✗	Identifying specific information	News article	Animal ethics
✗	Understanding main ideas and supporting ideas	News article	Animal ethics
✗	Understanding main ideas and supporting ideas	Academic article	History and culture
✗	Interpreting a word or phrase as used by the writer	News article	Animal ethics
			DELTA 113
✗	Identifying specific information	Fiction	History and culture
✓	Understanding main ideas and supporting ideas	News article	Animal ethics
✓	Understanding information and making an inference	Academic article	History and culture
✓	Identifying specific information	News article	Environmental issue
✓	Understanding information and making an inference	Academic article	History and culture
✗	Identifying specific information	Academic article	History and culture
✓	Identifying specific information	News article	Environmental issue
✓	Interpreting a word or phrase as used by the writer	Academic article	History and culture
✗	Identifying specific information	Fiction	History and culture
✓	Interpret a word or phrase as used by the writer	News article	Environmental issue
✗	Inferring the writer's reasoning	Fiction	History and culture
✓	Understanding main ideas and supporting ideas	Fiction	History and culture
✓	identifying specific information	News article	Environmental issue
✓	Understanding information and making an inference	Fiction	History and culture
✓	Interpreting a word or phrase as used by the writer	News article	Environmental issue
✓	Identifying specific information	News article	Environmental issue

Figure 2. Reading subskills tested in the DELTA diagnostic test.

details". The Diagnostic English Language Needs Assessment (DELNA) reading test assesses the following eight abilities: "ability to find specific information"; "ability to locate causes and effects, sequences, contrasts"; "ability to distinguish between main points and evidence or supporting ideas"; "ability to select words which fit the meaning and the grammatical construction of the text"; "ability to summarize main topics"; "ability to draw a conclusion based on information in a passage"; "ability to distinguish between fact and opinion"; "ability to organize information in a passage in another way (e.g. insert in chart, graph, map, diagram)" (www.delna.aucklang.ac.nz/uoa/).

In contrast, the Diagnostic English Language Tracking Assessment (DELTA) tests a range of subskills across various text types addressing a range of different topic areas (http://gslpa.polyu.edu.hk/eng/delta_web/). Figure 2 presents the reading subskills tested in DELTA, and gives an example of which subskills a particular learner performed poorly on (indicated with a "x" and highlighted). The diagnostic report is "based on the item difficulty relative to the student's proficiency (e.g., DELTA Measure 113 in this case). In other words, items that are of a lower difficulty level than the student's proficiency are those that the student would be expected to answer correctly. If they are not answered correctly, they indicate possible weakness in that particular subskill" (Urmston, Raquel, & Tsang, 2013, p. 74).

Although such subskills are frequently tested in reading research in both L1 and SFL, it is unclear to what extent they can be said to be diagnostic. First, it is almost certainly the case that such subskills do not exist in isolation but are involved in various combinations in reading any particular text or even answering any particular question. It is highly unlikely that understanding cause and effect, for example, does not involve the ability to

make inferences. Brunfaut and McCray (2015) found instances in which L2 readers prioritized and were able to make use of higher-level skills, but ignored or did not make use of syntactic information in the text (despite showing evidence of effective syntactic parsing in other instances) and did not arrive at a correct understanding, whereas successful readers showed that the integration of several processes was crucial. Thus the attempt to isolate subskills into discrete items seems doomed. Indeed, research (Alderson, 1990a, 1990b; Alderson & Lukmani, 1989) has shown that expert judges are unable to agree on what subskills are tested by individual items, and that there is no obvious hierarchy of difficulty in the relationship among subskills, such that, for example, one cannot make inferences before understanding specific detail, or understanding the main idea, or evaluating a text. Such levels of understanding depend very much both on the genre and topic of the text and the knowledge and reading purpose of the reader. Empirical research such as McCray, Alderson and Brunfaut (2012), which used a combination of eye-tracking and post-task interviews, has furthermore demonstrated that a diverse range of processes is used by test-takers who demonstrate similar understandings of a text and that differences in the employment of processes tend to be associated with differences in test-takers' proficiency level.

What seems clear is that fluent, higher-level reading depends on working memory capacity, attention and the automaticity of word recognition. In other words, such cognitive processes need to be mastered before, or at the same time as, the higher levels of processing are engaged. Thus diagnosing reading at the lower level is necessary (see, e.g., the list of cognitive measures for this purpose used by DIALUKI and referred to above). However, quite how to diagnose problems at the higher level, or problems related to the interactions between lower- and higher-level processes, is less clear.

Linguistic knowledge and diagnosing SFL reading

What SFL reading researchers often do is to assess a learner's linguistic knowledge and use that to predict one's reading ability. It is well known, for example, that having a small vocabulary in the SFL inhibits comprehension and conversely having a large vocabulary correlates highly with reading comprehension (Nation, 2001; Read, 2000). In addition, familiarity with less frequent words has a beneficial impact on one's ability to understand text. Thus, although Hughes (2003) claims that "the usefulness (and indeed the feasibility) of a general diagnostic test of vocabulary is not readily apparent … we would not normally require, or expect, a particular set of lexical items to be a prerequisite for a particular language class" (p. 179), we would argue that a measure of one's vocabulary size and depth would be very useful, especially if, as with the versions of the Vocabulary Levels Test studied in Schmitt, Schmitt and Clapham (2001), results were reported in bands of the frequency levels of the occurrence of words.

Assessing a learner's knowledge of grammar is somewhat more problematic, because the learning difficulties are likely to relate to one's first language, at least in part, and therefore one would need a diagnostic grammar test for each L1. Moreover, as Hughes (1989) suggests, a truly diagnostic test,

would need a number of examples of the choice the student made between two structures in every different context which we thought was significantly different and important enough to warrant obtaining information on. A single example of each would not be enough … As a result a comprehensive diagnostic test of English grammar would be vast. (p. 13)

One solution to this dilemma was developed by Johns' Birmingham Assessment and Diagnostic Test (BADT, 1971), the first version of which contained 130 items. The criteria for inclusion of items in the test were that they were: "important in written academic discourse; problematic for learners of English from a wide range of L1 backgrounds and levels of competence; covering areas which are not normally taught explicitly in English courses or which are badly taught" (p. 1).

An alternative solution, adopted by DIALUKI, was to develop a segmentation test, where all punctuation marks, formatting and spaces are removed from a text and the learner has to indicate where each word ends. For example, the learner is presented with the following text,

sothenextdaythethreelittlepigslefthomethefirstpigmadeahomefromstrawthesecondpig madeahomefromsticksbutthethirdpigwascleverhemadehishomefrombricksonedaythebig badwolfcametothestrawhouseheknockedonthedoor

and has to produce the following:

so|the|next|day|the|three|little|pigs|left|home|the|first|pig|made|a|home|from|straw|the|second| pig|made|a|home|from|sticks|but|the|third|pig|was|clever|he|made|his|home|from|bricks|one|day| the|big|bad|wolf|came|to|the|straw|house|he|knocked|on|the|door

In fact, such segmentation tests were used in the DIALUKI project as an indicator of parsing ability in both L1 Finnish and FL English and the results correlated significantly but only weakly or moderately with L1 but strongly with FL reading ability (see Table 1). The segmentation task turned out to be almost as good an indicator of reading in English as the more global measure of English vocabulary knowledge. The strong relationship of reading with the segmentation task may be due to its integration of vocabulary and grammatical knowledge (and it is worth noting that many linguists deny the dichotomy of grammar and vocabulary and prefer to refer to a continuum of lexicogrammar – see Alderson & Kremmel, 2013). However, it is difficult to say how "diagnostic" such segmentation tests are, and it is likely that several different abilities are involved in successfully completing the test, including word and morpheme recognition, syntax, knowledge of vocabulary and possibly sentence structure. A more focused test of syntax is probably preferable, in so far as it is possible to predict what aspects of syntax and morphology might be useful for reading at particular levels or ages or with learners of a particular L1.

It has been claimed (e.g., Alderson, 2005) that the more focused the diagnosis, the more useful. This is sometimes known as the grain size problem. The question is what is the optimal grain size for a particular diagnosis? Might this depend on the purpose of reading? Might it vary with the linguistic feature or cognitive variable to be targeted? For example, it is conceivable that diagnosing a learner's problems with modality would require a much more fine-grained diagnosis than problems with the passive, or exploring

Table 1. Correlations with L1 and SFL reading (Spearman rho).

	Rho with L1 Reading		Rho with SFL English Reading	
	L1 segmentation	L1 vocabulary	SFL segmentation	SFL vocabulary
4th grade	.27 (*n* = 190)	.34 (*n* = 193)	.55 (*n* = 194)	.59 (*n* = 203)
8th grade	.17 (*n* = 191)	.37 (*n* = 190)	.60 (*n* = 191)	.69 (*n* = 191)
Upper secondary	.21 (*n* = 188)	.23 (*n* = 195)	.53 (*n* = 172)	.64 (*n* = 171)

working memory compared with establishing automaticity of word recognition. A related matter is how many items are needed for an adequately reliable diagnosis? Those researchers retrofitting a diagnostic function to proficiency tests have experienced the problem that there are often simply not enough items addressing a particular feature for reliable results (Jang, 2009; Li, 2011).

Diagnostic feedback

Feedback to the learner and the teacher is an essential element in diagnostic testing, and a range of different types of feedback is possible, as discussed by Hattie and Timperley (2007) and Alderson et al. (2015). Feedback needs to be varied and avoid an over-emphasis on correctness, which may not help learners to understand the nature of their weaknesses or how to perform better in future. Feedback that helps learners to understand their own learning and reading strategies would be particularly useful. Feedback that encourages learners to monitor their comprehension (see, e.g., Purpura, 1998), or that provides hints and clues as to possible actions learners might take to reach a correct answer, could be of value. Computer-based diagnostic systems like DELTA and DIALANG offer generic advisory feedback as to what learners might do to improve in particular areas. However, at present, computers are limited to providing rather general and pre-programmed feedback. For example, the sort of feedback seen in Figure 2 simply tells learners in which subskills they appear to have problems but not what the underlying problems are. Research into how computers might be able to offer individualized feedback to learners on their specific performance, and fine-grained advice on how to improve, would certainly be welcome.

Computer-based testing is particularly useful for giving immediate feedback to learners on the results of the diagnosis. Feedback may be more useful if it is given when the experience of taking the test is fresh in the learner's mind, especially when the feedback relates to errors in task performance (Clariana, Wagner, & Roher Murphy, 2000), but learners have been found to be divided as to whether they like receiving feedback immediately after they have responded to an item, or whether they would rather wait until the end of the test (Yang, 2003). Computer-based testing makes it possible for learners to decide for themselves whether they receive immediate feedback or feedback delayed until the end of the test (a facility offered, e.g., in DIALANG).

Dynamic Assessment (DA – Poehner & Lantolf, 2013) is an interesting recent development which involves interaction between a teacher and a learner during an assessment

in order to assess a learner's potential for development. The teacher intervenes during a learner's performance in order to ask leading questions or give hints in a process known as 'mediation'. Recent attempts have been made to computerize such mediation; for example, the Computerized Dynamic Assessment of Language Proficiency (CODA), which is said to provide more fine-grained profiles of listening and reading, including how many attempts the student made, which help or support they made use of, and what effect such support had. Items are accompanied by pre-scripted prompts, triggered by incorrect responses, in order to give clues as to how the learner might improve on a particular item. At present, the nature of these prompts seems fairly rudimentary (such as "read the second paragraph again" or "pay attention to the punctuation") but research into what graduated adaptive feedback and hints learners find useful should lead to refinements to the suggestions that the system can offer in the future.

Treatment

An interesting feature of current development in the DELTA diagnostic system is the linking of feedback to "DELTA-dedicated learning resources, such as teacher/advisor support and self-access learning provision, both physical and online. These resources are gradually being put in place … through initiatives such as the Inter-university Collaborative Online Self-Access (ICOSA) project, which is developing an online repository of self-access English language learning resources for Hong Kong university students" (Urmston, Raquel, & Tsang, 2013, pp.77–78). DIALUKI is experimenting with two different interventions, one aiming to increase learners' morphological awareness in both the L1 (Finnish) and the SFL (English), and the other exploring vocabulary learning strategies and extensive reading. Pre-tests, post-tests and delayed post-tests are administered, as well as classroom observations, teacher reports and interviews, and interviews with focal students.

Issues arising include for how long such interventions should last – a month? a year? – and how many class hours per week can be spared for such interventions in addition to or instead of normal teaching following the regular syllabus. In addition to such practical issues, however, is the more general and theoretical issue as to whether the success or failure of such interventions invalidates the initial diagnosis.

Although diagnosis and treatment are separate areas, diagnosis is intended to lead to treatment, and the more specific the diagnosis can be, the more likely it is that useful teaching and learning materials can be devised.

Diagnostic assessment of SFL listening

Listening has been one of the most under-researched aspects of assessment, reflecting its "Cinderella" status among the four skills in second language learning and teaching research more generally. Within the research that does exist in the field of testing and assessment, the diagnostic assessment of listening is particularly under-represented. There have been several studies which have examined the diagnosis of listening post-hoc through the use of statistical models (e.g., Buck & Tatsuoka, 1997; Lee & Sawaki, 2009), and Alderson (2005) has written about the development of the listening section of

DIALANG. Despite very different approaches to diagnosis in general, these studies have shared a similar view of how diagnosis of listening might be "done" for SFL listening, and this is to measure, and report performance on, discrete subskills associated with componential views of listening proficiency. This may not be a tenable approach, however, as the reality of subskills is yet to be agreed upon (Alderson, 2000; Buck, 2001). Further, there is little practical use in diagnosing according to a subskills approach if diagnosis is theorized (as we have suggested) as a precursor for future intervention. Of what practical use is it for a teacher to know that a student needs more help with "listening for general ideas"? It is likely that diagnostic information of this kind – which arguably describes a further symptom rather than the actual cause of a comprehension problem – will result in the type of activities which are common in the SFL listening classroom: administering tasks in a test-like environment which simply "practise" a particular subskill (see Sheerin, 1987).

For diagnostic purposes (according to the principles outlined at the beginning of this paper), a more fine-grained approach is required if useful diagnostic tools are to be designed, and if follow-up treatment is to be recommended for addressing particular listening difficulties. One option would be to consider lists of "micro-skills" such as those compiled by Richards (1983), which provide a view of listening at a smaller grain-size than the broader subskills (main ideas, gist, specific details) usually associated with diagnostic listening assessment.[3] Richards' (1983) lists of micro-skills, which have been re-presented elsewhere (most notably in Buck, 2001), are divided into micro-skills for "conversational listening", and micro-skills for "academic listening". Some examples of micro-skills for conversational listening are as follows:

- Ability to retain chunks of language of different lengths for short periods
- Ability to discriminate between the distinctive sounds of the target language
- Ability to detect meanings expressed in different grammatical forms/sentence types
- Ability to process speech at different rates
- Ability to distinguish between literal and implied meanings

(Richards, 1983, pp. 228–229)

The micro-skills are a crucial feature in Richards (1983) own recommended "Design" process for planning a listening curriculum, which includes, first, a needs analysis, then the selection of micro-skills according to the types of listening activities identified by the needs analysis, then diagnostic testing to determine learners' current strengths and weaknesses with respect to these micro-skills, and finally, the formulation of instructional objectives based on these results. In short, Richards suggests a curriculum design process which reflects some of the stages of our own diagnostic assessment process, although in his case the process serves to identify objectives for a whole class, while in our case it is highly individualized and would, ideally, result in work-plans for individual learners.

However, the micro-skills approach has limitations as a basis for diagnostic assessment. Richards does not explicitly suggest how micro-skills might be operationalized, and in fact seems to suggest a "catch-all" diagnostic/proficiency test which could yield information about selected micro-skills. Another limitation of the micro-skills lists is

their atomistic nature; there is no unifying principle to the list, other than a single broad context for listening. Without a unifying principle, the process of diagnosis becomes atheoretical. The diagnostician, in other words, might have an understanding of a particular problem and its remedy, but not how or where that problem is located within a whole, fully functioning system.

It is likely to be more helpful, then, for the purposes of designing useful diagnostic tools, to turn to one of the process models of listening comprehension which have been proposed (e.g., Anderson, 2009; Rost, 2011). Such models theorize the various levels of processing required in comprehending natural speech. It is important to note that they also provide a framework for understanding the point of breakdown in the listening comprehension process as a whole. One useful example is the recently proposed model of SFL listening proficiency which has been articulated by Field (2013). This model contains five levels: input decoding, lexical search and parsing (which are described as lower-level processes), and meaning construction and discourse representation (described as higher-level processes). Importantly, comprehension does not follow a strictly linear progression from the lower to the higher processing levels; rather, different levels may be operating concurrently, with breakdowns at one level compensated by "positive information" at another, or with simultaneous breakdowns at higher- and lower-levels leading to miscomprehension altogether. In this sense, we take the term "level" to be a useful metaphor for understanding the listening process rather than an actual fixed stage in a linear system. Field (2013, p. 97) elaborates on the model by including the following descriptions of "levels", "knowledge sources" and lower level processes of listening comprehension:

Input decoding [supported by phonological knowledge]

- Phonemic decoding
- Syllable-level decoding
- Suprasegmental information

Lexical search [supported by lexical knowledge]:

- Phonological match
- Segmentation cues
- Word frequency
- Spreading activation

Parsing [supported by syntactic knowledge]

- Syntactic parsing
- Word sense narrowed
- Intonation

(Field, 2013, p. 97)

In Field's model, by the parsing stage, the incoming input will have moved from phonological string, to word string, to proposition form – a literal understanding of the

speaker's words. At the higher-levels of processing, this proposition is then related to the "circumstances in which it was produced" (Field, 2013, p. 100) in order to extract meaning in the "meaning construction" phase. These "circumstances" include the context of the utterance, as well as the listener's understanding of the speaker's intentions. The onus is then on the listener to draw on various knowledge sources – pragmatic knowledge, external knowledge (e.g., knowledge of the world), and discourse representation (the listener's knowledge of the text/interaction up to that point) – in order to make sense of the utterance. Field's highest level is "discourse representation", and at this point the listener integrates what they have understood from an utterance into an understanding of the whole listening event. This involves the selection of relevant material to integrate into a developing macro-structure, and draws again on external knowledge as well as local knowledge of, for example, text type.

There is obvious scope, here, for operationalizing elements of this model through discrete assessment tasks. As with reading (see above), this is likely to be much more straightforward at the lower-levels of processing where the influence of information beyond the text – such as context – and knowledge sources which are not traditionally labelled as "linguistic" is much less pronounced (although the interactive nature of the model means that context is, in fact, always an influence). However, there is also scope for an expansive set of diagnostic tools which take into account, too, the strengths and weaknesses in various linguistic and extra-linguistic knowledge sources that are crucial for meaning-making and for building a structure of a listening event. Some potential test types related to each level of processing – which would need to be prototyped and trialled more fully – are suggested in Table 2:

Following Field's model, it is important to note that in some cases a listener may experience difficulty at a specific level of processing, and at other times the problem might be caused by faulty interactions between different levels of processing working together. For this reason, ideally there should be an additional layer of diagnosis targeting process interactions, though how this might be measured in practice would warrant research.

A worked example

In order to illustrate how the use of tools of this kind could be utilized within the diagnostic process (see the five principles, and specifically Principle 4), we will now provide a worked example of how a particular listening problem might be further investigated within the diagnostic process as outlined in Figure 1.

Context. We can imagine a hypothetical situation where a teacher is interested in finding out about the strengths and weaknesses of a group of adult learners in an English for Academic Purposes (EAP) listening class. Following the diagnostic process, the teacher/ diagnostician may take the following steps.

1. Listening/observing stage. At this stage, the teacher might have developed a detailed understanding about particular students already by observing learners' listening performance during classroom tasks, or by informal assessments of learners' listening ability during conversations. However, in order to make a more formal diagnostic assessment

Table 2. Potential tools for diagnostic listening assessment.

Level of processing	Discrete listening assessment tasks for diagnostic purposes
Input decoding	• Phoneme discrimination task (utilizing diverse speakers) • Counting syllables in familiar/unfamiliar words • Identifying stressed syllables in a word (e.g., underlining syllables) • Identifying stressed words in a sentence (e.g., underlining words) • Identifying intonation patterns (e.g., choosing from line representations of intonation patterns)
Lexical search	• Word recognition task (e.g., repeat last word in a sentence) • Counting the number of words in an utterance • Aural word association tasks • Aural vocabulary tests
Parsing	• Disambiguating "garden-path" sentences (delivered aurally with neutral tone/prosody) • True/false judgements of literal utterances focusing on particular elements of spoken syntax
Meaning representation	• Aural tests of pragmatic knowledge, including comprehension of speech acts and implicature as well as sociopragmatic knowledge • Aural tests of metaphor comprehension • Tests of implicit information (e.g., inferential knowledge) • Tests of "social listening" knowledge (e.g., recognition of listening events/texts/genres)
Discourse construction	• Distinguishing between relevant and irrelevant information in a lengthy text (when listening for a specific purpose) • Storing information heard earlier in a text for review and modification according to information later in a text

Note: Speed of input in all tasks might be graded according to level. Processing speech automatically in real time is a crucial factor in SFL listening comprehension, and this variable, along with other speaker-related features (pronunciation, fluency, age, etc.) would need to be modelled in task input according to need.

about a listener's ability to follow a lecture, following the diagnostic process the teacher has a number of choices:

1. The teacher may ask a student to self-assess the strengths and weaknesses of his/her listening ability against a checklist, itself based on a process model of listening comprehension (this may be produced in the L1 or the SFL). In this way, the teacher gains an emic view of the learner's listening ability.
2. The teacher may choose to administer a level-appropriate general listening task (e.g., a task that focuses on comprehension of main ideas and specific details, or a task which focuses on understanding implicit meaning in a conversation), and then discuss with individual students afterwards the type of difficulties they experienced during the comprehension process. This sort of teacher–learner

conference would have many of the hallmarks of a stimulated recall procedure which is commonly used in research on SFL listening (see, e.g., Révész & Brunfaut, 2013).

For this example, we will assume that a teacher noticed a learner had been experiencing difficulty during classroom listening tasks. The teacher decided to choose option 2, and interviewed the student following the administration of a listening test (using the student's responses to the listening test as a stimulus for recollection of listening difficulty). In this hypothetical situation, during the course of the post-task interview, the teacher records several examples of reported listening problems for which the following exchange is typical:

Example of extract from post-task interview.[4]

Teacher: Okay um what answer did you have there for question twenty?
Learner: I think it's the improvement of … something (laughs) I missed that word, I can't follow what she's saying that words
Teacher: What ah what difficulty do you think you were having there?
Learner: I think she said that word so fast and I ca(n't) I can't record … for myself yeah …
Teacher: Okay …

2. Initial assessment. Based on the results of the interview, the teacher's knowledge of previous students who have experienced similar difficulty, and his/her understanding of SFL listening processes, the teacher hypothesizes that this particular student is finding it difficult to identify specific words in the stream of speech, suggesting a word identification problem, perhaps related to a lexical segmentation problem (given the reference to speed of speech), both of which fall under Field's (2013) category of lexical search. What is interesting about the example above is that the listener shows no particular problem in parsing the utterance she has heard. She is, in fact, able to construct the proposition despite not recognizing the "missing" word: "it's the improvement of … something". In this case, then, the teacher/diagnostician may develop a hypothesis that the listener has particular difficulty at the lexical search level when listening to authentic speech in real time.

3. Hypothesis checking. Ideally, at this point the teacher/diagnostician would have access to a repository of tasks – accessible online, comprising the range of possible tasks represented in Table 2 – which would help to confirm this hypothesis. If this were the case, the teacher/diagnostician could then choose to administer a word recognition task, for example where the learner repeats the final word they hear in a sentence or phrase. Given the problems associated with speed of delivery, this task might be graded by speech rate, beginning slowly and advancing in pace as the task progresses until the teacher/diagnostician can identify the rate at which the learner begins to find the task noticeably challenging. Alternatively, a task targeting lexical segmentation ability, such as the ability to count the number of words in a sentence (also perhaps graded by speech

rate) would provide evidence of the learner's ability to use phonological cues to isolate words in the speech stream. The results of these tasks would be used to provide evidence either to support or refute the initial assessment.

4 Decision making. In the final stage of the diagnostic process, the teacher/diagnostician would need to decide, first, whether or not the initial assessment has been supported by the evidence. If this is not the case – for example, if the learner in the hypothetical situation performed very well on these targeted assessment tasks – the teacher may need to reconsider the initial assessment. For example, the problem here might have been a lack of vocabulary knowledge (i.e., the learner simply did not know the word, and therefore could not recognize it, or even reproduce it in phonetic form because of the demands of concurrently parsing the sentence in real-time). In this case, the teacher/diagnostician may administer another test, or return to stage 1 and seek additional information from the learner.

If, however, the diagnostic assessment task does reveal a noticeable problem for a given learner, results on the task should be used as the basis for detailed feedback and follow-up support. Feedback could, again, be in the form of a teacher–learner conference where the meaning of the learner's performance on a particular discrete listening task is explained clearly. The type of "treatment" would depend on the nature of the classroom (i.e., class size, whether individualized work plans are feasible, whether students have a self-access centre), however further support should be guided by the particular difficulty identified by the diagnostic process. In the case of the hypothetical example, targeted treatment would *not* consist of, for example, extensive listening (although this might be helpful as an auxiliary approach) or suggestions to complete more textbook listening exercises. Rather, treatment would be intrinsically connected to word recognition and segmentation, with follow-up activities such as tailored dictation or gap-filling exercises suggested, optimally in a computer-based environment where learners can also control the rate of speech.

It is worth noting that the diagnostic process might be re-ordered, in some respects, depending on the type of initial assessment which is taken at the listening/observation stage. For example, a teacher might engage in dynamic assessment techniques, which according to Lantolf and Poehner (2008) can "bring to the surface processes that underlie performance and that often remain hidden during traditional assessments" (p. 276). If, through this process of mediation, a diagnostician is able to identify particular strengths and weaknesses in a processing capacity then it is feasible to move straight to decision-making (stage 4). In many respects, a well-devised dynamic assessment activity actually contains elements of all three initial stages – listening/observation; initial assessment; hypothesis checking – and leads to the appropriate type of decision-making within this paradigm.

Conclusion

Alderson et al. (2014) pointed out that the SFL testing field is in need of a more detailed theory which can account for the multifaceted nature of diagnosis in language assessment. To provide such guidance Alderson et al. (2014) explored the process of diagnosis

in other professional and vocational domains, and sketched a framework for diagnostic assessment as well as developing a set of five principles which could facilitate an exploration of the various facets of this framework with a view to improving diagnostic assessment practices more generally.

Starting off from these five tentative principles, this article explored and suggested practical answers to how diagnostic assessment may take place in the context of SFL reading and listening. A four-step diagnostic process derived from and encompassing the five principles was proposed (see Figure 1) and illustrated with a concrete example for SFL listening. At the heart of this diagnostic process lie insights into the constructs – in this case SFL reading and listening. Existing SFL reading diagnostic tests were critically considered and a range of discrete tasks were put forward for SFL listening diagnosis, whilst acknowledging the need for further research into tools aiming to diagnose so-called higher-level skills. Such instruments are necessary to extend our diagnostic assessment capacities as related to semantic and pragmatic aspects of the SFL reading and listening constructs. Importantly, this will also help ensure capturing context-related dimensions of language proficiency.

The proposed, fine-grained diagnostic approach should in principle be compatible with any sort of syllabus, since it focusses on those processes that make communicative performance possible. Follow-up to the diagnosis, however, will not solely need to be based on the specific diagnosis; effective treatment also requires alignment with the principles of the specific curriculum in which the diagnosis is taking place. For example, treatment in the context of a task-based language teaching pedagogy will need to identify a way to merge discrete diagnostic information with a holistic, task-outcome-focused approach. However, the need for a symbiotic relationship between curriculum, diagnostic assessment and treatment requires that theories of language and of learning that are embedded in classroom contexts are first made explicit, which may not always be the case. Research exploring successful translations of diagnosis into treatments across a range of pedagogic contexts would thus be desirable.

Funding

Notes

1. For a complementary analysis of the same data set, see Alderson et al. (2015).
2. Please note, within this figure we have assumed the teacher would take the role of diagnostician. This is just one option, however, and diagnosticians may also be an expert diagnostician, an applied linguistics/second language acquisition expert, learners' peers, learners themselves, an adaptive computer-based diagnostic system or any combination of these working together.
3. Although note that other micro-skills in the taxonomy seem quite broad, for example, "Ability to identify topic of lecture and follow topic development" in the academic listening list.
4. Note: this transcript is drawn from authentic data which formed part of a dataset collected for a project reported in Harding (2011). The data from this particular learner was eventually not used in that study, but provides a useful example here of a learner commenting on her own listening difficulties.

References

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds), *Reading in a foreign language* (pp. 1–24). London: Longman.

Alderson, J. C. (1990a). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language*, *7*, 465–503.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part One). *Reading in a Foreign Language*, *6*, 425–438.

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

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York, NY: Continuum.

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*. doi: 10.1093/applin/amt046

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validity of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *34*, 535–556.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*, 253–270.

Anderson, J. R. (2009). *Cognitive psychology and its implications* (7th ed.). New York: Worth.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. (ARAGs Research Reports – Online. Vol. 1, No. 1). London: British Council.

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*, 119–157.

Carver, R. P. (1985). How good are some of the world's best readers? *Reading Research Quarterly*, *20*, 389–419.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*, 84–95.

Carver, R. P. (2000). *The cause of high and low reading achievement*. Mahwah, NJ: Lawrence Erlbaum.

Clariana, R. B., Wagner, D., & Roher Murphy, L. C. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, *48*, 5–22.

Clay, M. (1985). *Reading: The patterning of complex behavior* (3rd ed.). Portsmouth, N.H.: Heinemann.

DELNA (Diagnostic English Language Needs Assessment). Retrieved December 31, 2013 from www.delna.auckland.ac.nz/uoa/

DELTA (Diagnostic English Language Tracking Assessment) Retrieved December 31, 2013 from http://gslpa.polyu.edu.hk/eng/delta_web/

DIALANG Retrieved December 31, 2013 from www.lancaster.ac.uk/researchenterprise/dialang/about.htm/

DIALUKI (Toisen tai vieraan kielen lukemisen ja kirjoittamisen diagnosointi – Diagnosing reading and writing in a second or foreign language). Retrieved December 31, 2013 from www.jyu.fi/dialuki

Ehri, L. C., & Snowling, M. J. (2004). Developmental variation in word recognition. In C. A. Stone, E. R. Silliman, B. J. Ehren & K. Apel (Eds.), *Handbook of language and literacy. Development and disorders* (pp. 433–460). New York: Guilford Press.

Field, J. (1998). Skills and strategies: Towards a new methodology for listening. *ELT Journal*, *52*(2), 110–118.

Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds). *Examining listening: Research and practice in assessing second language* (pp. 77–151). Cambridge: Cambridge University Press.

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

Guthrie, J. T., & Wigfield, A. (1997). *Reading engagement: Motivating readers through integrated instruction*. Newark: International Reading Association.

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. (Language Testing and Evaluation). Frankfurt: Peter Lang.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112.

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

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, *26*(1), 31–73. doi: 10.1177/0265532208097336

Johns, T. (n.d.). *Birmingham Assessment and Diagnostic Test*. Undated manuscript. Birmingham, UK: Birmingham University.

Johns, T. (1976). *Why diagnose?* Memorandum dated July 18, 1976. Birmingham, UK: Department of English, University of Birmingham.

Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Kintsch, W. (1998). *Comprehension: A framework for cognition*. New York: Cambridge University Press.

Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.

Lantolf, J. P., & Poehner, M. E. (2008). *Sociocultural theory and the teaching of second languages*. London: Equinox.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*, 239–263.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (Vol. 9, pp. 17–46). Ann Arbor, MI: University of Michigan.

McCray, G., Alderson, J.C., & Brunfaut, T. (2012, June). *Combining eye-tracking with post test interview data to examine gap-fill items: Triangulation of tribulation*. Paper presented at the European Association for Language Testing and Assessment conference (EALTA), Innsbruck, Austria.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Perfetti, C. (1985). *Reading ability*. New York: Oxford University Press.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research*, *17*, 323–342.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modelling approach. *Language Testing*, *15*(3), 333–379.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*, 31–65.

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240.

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Abingdon, Oxon: Routledge.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*, 55–88.

Sheerin, S. (1987). Listening comprehension: teaching or testing? *ELT Journal*, *41*(2), 126–131.

Snowling, M. J., & Hulme, C. (2005). *The science of reading. A handbook*. Malden, MA: Blackwell.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics*, *14*, 60–82.

Vandergrift, L., & Goh, C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.

Yang, R. (2003). Investigating how test-takers use the DIALANG feedback. (Unpublished master's thesis). Lancaster University, Lancaster, UK.