5 The effect of memorized learning on the writing scores of Chinese IELTS test-takers

Authors

Alison Wray
Cardiff University, UK

Christine Pegg
Cardiff University, UK

Grant awarded Round 11, 2005

Presents a method for establishing the proportion of potentially memorized material in the performance of IELTS candidates in the academic writing task 2.

ABSTRACT

We address the challenge of assessing performance when IELTS (Academic Writing Task 2) candidates may have memorized and reproduced lengthy chunks of text that potentially disguise their true proficiency. Our profiling procedure separates out text that is more and less likely to reflect the candidate's genuine linguistic knowledge. The procedure was applied to 233 retired scripts by Chinese candidates, and the results are analyzed by band and test centre.

As expected, errors decreased as band increased. Similarly, the quantity of non-generic nativelike text increased with band. But the use of material copied from the question and of 'generic' nativelike text (text that can be used in most essays) remained constant across bands for all but one test centre. Using the mean profiles as norms, a script known to be problematic was examined, to demonstrate how profiling can isolate the nature of differences. Three less extreme 'outlier' scripts from the main sample were also examined, to help locate a threshold for what counts as a problem, and demonstrate why unusual profiles can occur. To assist examiners, a simplified version of the profiling procedure is offered, that can be used as an informal diagnostic.

The profiling procedure recognizes the legitimacy of producing some pre-memorized nativelike material in a writing test, by contextualizing it within the broader pattern of the candidate's written performance overall. The procedure requires further refinement than was possible within this modest project, but already suggests potential strategies for IELTS examiners to recognize memorized material in writing tests.

AUTHOR BIODATA

ALISON WRAY

Alison Wray is a Research Professor of Language and Communication in Cardiff University's School of English, Communication and Philosophy. Her research activity has focussed on the processing and interactional functions of formulaic language in normal, abnormal and learner discourse; the evolution of language; and language profiling for applied purposes. Her 2002 monograph Formulaic Language and the Lexicon (Cambridge University Press) was awarded the 2003 book prize of the British Association for Applied Linguistics. Another book, Formulaic Language: Pushing the Boundaries, was published in 2008 (Oxford University Press). She has also co-authored two highly successful textbooks, Projects in Linguistics (Hodder Arnold, with Aileen Bloomer) and Critical Reading and Writing for Postgraduates (Sage, with Mike Wallace).

CHRISTINE PEGG

Christine Pegg is a Lecturer in the Centre for Language and Communication Research at Cardiff University. She is a certificated IELTS Examiner, IELTS Examiner Trainer and the Examiner Support Coordinator for the IELTS Professional Support Network for UK and Ireland. She has conducted EFL Oral Examiner training in both Argentina and Cyprus, and delivered an intensive MA course on the teaching and testing of grammar in Caracas, Venezuela. Her present work focuses on language testing in China, where she is a Guest Professor at Tianjin University of Technology, Tianjin, and Lanzhou University. Her primary research interests are language testing, assessment and evaluation, TEFL teaching methodology and teacher education.

IELTS RESEARCH REPORTS, VOLUME 9, 2009

Published by: British Council and IELTS Australia
Project Managers: Jenny Holliday, British Council Jenny Osborne, IELTS Australia
Acknowledgements: Dr Lynda Taylor, University of Cambridge ESOL Examinations
Editor: Dr Paul Thompson, University of Reading, UK

ISBN 978-1-906438-51-7 © British Council 2009 Design Department/X299
The United Kingdom's international organisation for cultural relations and educational opportunities.
A registered charity: 209131 (England and Wales) SC037733 (Scotland)

CONTENTS

1 INTRODUCTION

How is a candidate in the IELTS test to convince the examiner that he or she should receive a high mark? The obvious answer is: by using the language proficiently. But what if the most proficient-*looking* language does not require the greatest proficiency to produce? Memorized linguistic material could constitute such a case. Although it is, of course, possible to have a full command of what one memorizes—as is the case with actors learning a script, for instance—there is clearly the potential to demonstrate, in the reproduction of memorized phrases or sentences, a level of linguistic sophistication beyond the reach of one's real productive competence.

Because of this possibility, judging a learner's proficiency on the basis of the amount of nativelike output is not a straightforward matter. While non-nativelike output can be taken as a reasonable gauge of proficiency limitations, nativelike output can be produced at many different stages of learning, and can signify many different things. A complete beginner could correctly write out a memorized sentence while an intermediate learner, trying to express the same idea from scratch, made errors. In certain contexts, then, nativelike output might even be judged as suspiciously *too* correct, and the temptation would be to mark it down. Yet an assessor has no way of knowing the provenance of nativelike material in the learning and production of a candidate, and therefore no way of distinguishing between those who use it to disguise their true ability and those who use it as a legitimate expression of that ability. The IELTS marking scheme rewards nativelike language, and cannot be expected to discriminate between the different possible motivations for its production.

Recent work in psycholinguistic theory and second language acquisition theory presents an additional complication to the picture. It has been proposed that the native speaker him/herself achieves idiomaticity through the memorization of useful wordstrings (Wray 1999, 2000, 2002a). Furthermore, there is substantial evidence that material so memorized may not, even in the native speaker, have been subject to the kind of analysis that in former theories was considered central to having a genuine 'command' of it. This 'formulaic language' appears to pervade nativelike performance, though estimates of its proportion in natural language vary from 4% to 80% (see Wray 2002a, pp 28ff for a discussion of why). For our present purposes, the higher figure is certainly too extreme to be useful. It includes a much broader range of linguistic configurations, including collocations, that we know the learner must genuinely master in order to gain advanced competence in the language. However, there is a subset of material that not only learners but also native speakers may very deliberately memorize as part of the development of a reliable exam technique or as part of their academic writing skills. Here, there would be a particular irony and unfairness, were the learner to be penalized for using memorized nativelike wordstrings for structuring an assessed essay, when the native speaker legitimately memorizes and employs them to the same ends.

If the learner who memorizes useful wordstrings is, in fact, emulating the native speaker, and if the outcome is communicatively apposite and grammatically accurate, there can really be no grounds at all for not awarding high marks. Yet, as noted, memorization may disguise relatively low levels of general command of the language, and it would be inappropriate to reward a learner for stringing together ill-understood material. While the 'joins' between memorized strings may reveal something of the true level of ability, the underlying problem remains—that of defining fairly and accurately for all candidates what should, in fact, be the most acceptable parameters of 'true level of ability'.

It follows that some distinction must be made between 'appropriate' and 'inappropriate' reliance on prefabricated linguistic material on the part of the IELTS candidate. The question, though, is how that can be done without undermining the robustness of the IELTS marking criteria.

The research project reported here has developed a practical means of *profiling* of a writing task response, so as to gauge its typicality to norms based on band score. Since examiners are generally very able to identify problematic scripts, there has been no need to develop an *ab initio* tool—the aim was not to replicate or challenge the efficacy of the existing assessment rubric. Rather, the opportunity is presented for an examiner to explore the basis for his or her disquiet about a given script, and to ascertain with relative speed the extent to which aspects of the profile, diverging from the norm, support the concerns about it.

The data used for this research are retired scripts from the IELTS writing task 2 (academic), all written by Chinese candidates. Chinese candidates were used because of the popular perception that the Chinese educational tradition favours rote learning. In fact, recent research demonstrates that the picture is much more complex. However, it does confirm that Chinese learners perceive a tangible value to memorization, provided it is accompanied by understanding (see Section 3). With the sharp rise in English language proficiency targets in China—China has been the top location for IELTS candidature since 2002—and the evident benefits for the individual with recognized qualifications in English, the consequences of a memorization tradition are being

seen in test performance, both oral and written. The fair and accurate assessment of Chinese candidates has therefore been perceived as a particular challenge.

2 AIMS OF THE PROJECT

The project had three key aims:

To investigate the effect of memorization on the writing task scripts (Academic, Task 2) of Chinese mother tongue IELTS candidates.

To develop a method for identifying candidates who may, in the IELTS writing task (Academic, Task 2), have used excessive amounts of memorized material, to the extent that it inflated their score.

To streamline the method to the point where it can be used by IELTS examiners as a diagnostic for suspect scripts, without the need for software or complicated calculations.

The analyses were focussed around the following research questions:

1 How can writing task responses be profiled to indicate potential levels of pre-memorization?
2 What is the relationship between profile features and band score?
3 Do candidates from different test centres display different profiles in relation to the amount of potentially memorized material they use?
4 Is it possible, on the basis of norm referencing by profile, to identify problems in a script?
5 What is the simplest profile measure that can be used as the basis of diagnosis?
6 Is it possible to place scripts on a continuum, in relation to less striking tendencies towards the overuse of memorized material?

3 CONTEXT

3.1 Memorization in the Chinese educational tradition

A number of recent studies review and explore the role of memorization for Chinese students (eg, Au and Entwistle 1999; Cooper 2004; Dahlin and Watkins 2000; Ding 2007; Kennedy 2002; Ting and Qi 2001; Zhanrong 2002). All seek to dispel the myth that memorization is confined to surface learning, and argue that, in fact, "differences in the role of memorization are at the heart of the commonly found superior performance of Asian compared to Western students" (Dahlin and Watkins 2000, p 66). The key to this association is the use of memorization to consolidate and/or facilitate understanding (*ibid*, p 67; Cooper 2004, p 294).

For Marton *et al* (1993, p 10) "Memorization with understanding' has two components: 'Memorizing what is understood' and 'Understanding through Memorization'. That is, memorization serves an end in itself (if you can't remember something, you cannot use it) and also enables 'the discovery of new meaning" (Dahlin and Watkins 2000, p 80). However, these observations relate to subject learning rather than language learning. The rote learning that goes on in vocational or academic subject areas such as Business and Accountancy (eg, Cooper 2004) entails the capacity to ensure that vital information is easily available. Thus, memorization becomes a means by which the human brain is used as a substitute for the notebook (paper or electronic) that is not permitted in the exam hall. The technique of cramming the head full of memorized facts, so that one has a database from which to select relevant material under pressure, is certainly not restricted to the Chinese tradition, but rather is an inevitable consequence of the testing process. But to what extent can this technique be used also to learn linguistic forms?

3.2 Language learning through memorization

The question just posed resonates with a major debate dating back several decades, which Wray (eg, 1999, 2002a) reviews in detail. It revolves around the extension into language learning of Marton et al's (1993) claims, namely: is the memorization of linguistic material (a) only possible if the make-up of it is fully understood, or (b) an opportunity to store now and analyze later, by creating a pool of linguistic material upon which the brain can work either subconsciously or consciously in the future? Wray's review suggests that both may apply. Since (b) seems to apply during first, and early childhood second language acquisition, there is some interest in establishing whether older second language learners also have the capacity for the 'learn first,

analyze later' approach, even if educational traditions and personal expectations tend to direct the learner towards a preference for (a).

Indications that L2 learning after early childhood may proceed on the basis of both (a) and (b) come from Ding (2007). Regarding (a), memorization founded on understanding, Ding notes that usage-based learning can particularly benefit from memorization strategies. During interaction, one is confronted by the shortfall between the input and one's capacity to produce adequate output, but there is little time either to notice the nature of the shortfall, or to consolidate noticed new forms through immediate rehearsal (p 272). Off-line memorization furnishes opportunities to bring a more systematic attention to forms, and to practise them, so that they are more easily put into use under the demands of real time communication. Furthermore, memorization delivers "a relatively good feel for English" (p 277) which makes it easier to notice and learn new features.

With regard to (b), Ding's (2007) work also offers some indicators. He interviewed three winners of a national English speaking competition in China. All had attended the same secondary school, at which memorization was particularly emphasized. Specifically, the students were expected to imitate recordings of native speakers reading texts, and, in the teacher's office, "[t]hey had to recite the text verbatim and in the same intonation patterns as they had heard on the tape. The teacher would criticize them if they failed to do so" (p 274). The pressure on students to achieve a high quality result was very great, with texts several pages long being memorized in senior classes (*ibid*). Tests and exams, by focussing on linguistic patterns (grammatical, collocational, phrasal), reinforced the importance of text memorization (*ibid*). Although few will deny that understanding would assist in this Herculean task, for such learners memorization had to continue even in the absence of it: one informant said "I had to listen to [a tape] many ... times before I could follow it" (Ding 2007, p 277). In all events, memorization would precede the learner's full productive command of the forms. Indeed, one may assume that that was the rationale for the teacher's approach: the expectation of a subsequent conscious or unconscious backfilling of competence, drawing on what was stored in memory (*ibid*, p 279). The mechanism by which such learning was ultimately consolidated was, again, usage. During class discussions, the memorized texts would become a productive resource, so that—as one of Ding's informants observed— "what had been memorized became our own language" (Ding 2007: 275), until, as Ding himself notes, "when they speak English, lines from movies often naturally pop out, making others think of their English as natural and fluent" (*ibid*).

3.3 Memorization and patterns of achievement

Memorization is not an easy option for the learner, and success seems to depend on the intensity of both a teacher's insistence and a learner's determination (Ding 2007, p 279). Thus, we should expect to see considerable variation in practice and outcomes. Whether or not it is possible to ascribe the Chinese memorization tradition to the heritage of Confucianism (see Kennedy 2002, p 431ff for a discussion of this), even in the context of national teaching curricula it must be recognized that certain differences are sure to exist—between rural and urban learners, individuals with greater or lesser aspirations to travel outside of China, and, of course, on the basis of individual learning styles, aptitude and motivation. Most marked in this regard is the potential for difference between the learning styles and learning successes of students in different Chinese-speaking contexts, including Taiwan, Hong Kong, mainland China, Malaysia and the many other countries worldwide in which Chinese speakers may take the IELTS test either on arrival or after some period of residency (see Section 5.2).

3.4 Memorization in the context of testing

As Ding's (2007) study clearly shows, one key motivation for students to apply themselves to the difficult challenge of memorizing texts was the awareness that they would later be tested on their knowledge. Initially, it was a matter of avoiding reports of poor performance reaching home (Ding 2007, p 276). Later, though, success in tests became a motivator for study, and in this regard one can infer the potential for some measure of washback into the teaching method. However, overall, memorization must probably be regarded in instrumental terms in relation to tests. According to Ho et al (1999, p 48), in the context of an examination or performance, "memorizing lines or already understood facts may be required to ensure success" (quoted in Kennedy 2002, p 433). In other words, however much a Chinese learner may believe in memorization as either the product of understanding or a way of deepening understanding, there is a pragmatism about test taking. If it is perceived that rote memorization, even without real understanding, can enhance test performance, then rote memorization will naturally become part of the preparation. This being so, it will ultimately be down to the testing bodies to respond. The difficulties inherent in doing so appropriately and fairly are a significant challenge to IELTS.

3.5 The native model: why memorization is authentic as well as effective

The theoretical rationale for the present research is that the memorization of multiword strings is a natural part of language learning, both for native and non-native speakers (Wray 2002a). On the basis of a detailed examination of evidence from first and second language acquisition, language loss, and patterns in discourse, Wray proposes that, in order to communicate effectively, humans use prefabricated, holistically stored, multiword strings in their output. These strings enable both the speaker and hearer to take processing shortcuts. The inventory of prefabricated strings contributes to characterizing the subset of grammatical material in a language that is also 'idiomatic'. The non-native speaker who, for whatever reason, does not store so much material holistically, is challenged to produce idiomatic forms by other means—that is, by constructing them out of smaller units by rule. This is both more effortful and, naturally, subject to potential overgeneralization and L1 interference. Even when proficient enough to avoid such errors, adult learners often produce output that is grammatical, meaningful yet not nativelike.

It follows that, logically, the goal for such a learner ought to be to match the native speaker's lexical inventory, by storing and retrieving the same large items. Recent investigations at Cardiff focus on whether this is in fact desirable, possible and effective (eg, Fitzpatrick and Wray 2006; Wray 2002b, 2004; Wray *et al* 2004; Wray and Fitzpatrick 2008; Wray and Staczek 2005). Findings so far indicate that there are considerable benefits for effective communication, but that adult learners find it very difficult to trust large units to memorization without fully understanding their form, and that once they do command the form, they tend to store the parts rather than the whole.

Thus, the relationship between memorization and understanding is complex, and the evaluation of idiomatic material in the output of a testee is going to be confounded by the following potential sequence in learning (Figure 1).

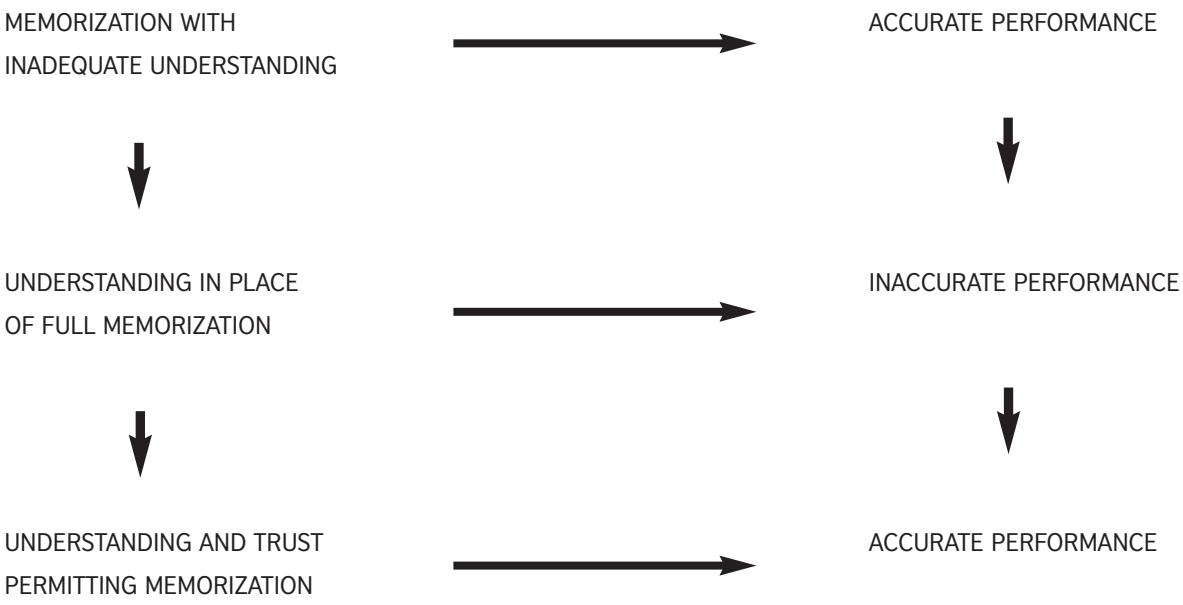


Figure 1: The progressive manifestation of accuracy in response to memorization

3.6 Assessing performance that includes memorized material

As Figure 1 indicates, an examiner is faced with a problem when assessing the accuracy of material that has been memorized. It is not only that accuracy may or may not disguise an absence of understanding, but also that inaccuracy may be indicative of greater understanding than some – but not all – accurate performance is. The examiner is charged, in short, with somehow differentiating between what might, in two candidates, be rather similar performances produced on the basis of considerably different ability. Clearly, certain common-sense considerations will apply:

1. An inadequately understood expression might be used inappropriately (though it might not).
2. One may judge the extent to which the material that is evidently *not* memorized is consistent with a particular band of ability.

Yet, in extreme cases, the first criterion may leave the examiner judging correctly used wordstrings not on their own merits but on the fact that, since some other wordstrings have been incorrectly used, the correctly used ones are likely to be lucky hits. Similarly, the second criterion may, in extreme cases, result in correctly formed multiword expressions being entirely ignored in favour of a judgement based only on the connecting material. Neither of these judgement strategies is desirable.

The diagnostic procedure developed in this project offers a means of distinguishing performances on the basis of a *profile* of the candidate's linguistic output. It has been framed in order to answer Research Question 1, *How can writing task responses be profiled to indicate potential levels of pre-memorization?* In what follows, the detailed profiling procedure is first described and evaluated. Then a streamlined version is presented, which draws its validity from the broader patterns of the detailed profiling. The streamlined version offers a means for examiners to operate with confidence and consistency in relation to this potentially very problematic material.

4 METHOD

4.1 Materials

This research is based on an analysis of IELTS Academic Writing Task 2 scripts. A general overview of the revised version of the Writing test (post January 2005) including the format, criteria and band descriptors is in Appendix 1 (taken from the IELTS Handbook 2007). The specific Academic Writing Task prompt used for this study is in Appendix 2.

Cambridge ESOL provided a total of 236 'retired' scripts (Academic Writing Task 2), all written by Chinese speakers. They had been allocated band scores between 2 and 9, but as there were only two scripts in Band 2 and only one in Band 9, these three scripts were not used. All the essays were responding to the same input prompt. The tests had been taken in IELTS centres in Australia (AU), Fiji (FJ), Hong Kong (HK), Malaysia (MY), New Zealand (NZ) and Taiwan (TW) (Table 1). The centres have been anonymised here, as AU (i) to (iv) etc.

4.2 Treatment

The scripts were transcribed into electronic text files and, following experimental profiling to develop the best approach, a set of criteria was drawn up for coding them (see below). Two native speakers were trained in the coding system. One coder was designated 'main coder' and she coded all of the data. The second coder coded a large subset of the same data for the purposes of reliability testing. The correlations between their judgements were highly significant (between .966** and .819**) for all but one profiling subtype (discussed below). These high correlations suggest that any native speaker following the criteria would reach somewhat similar subtype distributions to those of our coders. Maintaining and accepting the consistency of a single judge's subjective decisions, as opposed to combining and/or neutralizing the biases of two or more judges has its limitations (see later), but nevertheless most accurately reflects the likely application of this profiling technique, whereby a given IELTS examiner might sample for analysis a number of scripts for comparison with a problematic one.

The coding created a profile for each writing task response, and enabled the profiling of groups of writing task responses, such as by band and testing centre.

4.3 The profiling technique

The diagnostic tool offers a visual profile of a text that can show how it is constructed, specifically in relation to the balance between different key components that make a text nativelike and non-nativelike. The aim was to

minimize the focus on individual manifestations of nativelike and non-nativelike material per se, both because there is often more than one competing explanation for them (see earlier discussion), and because the existing approach to marking the scripts is assumed adequately to capture the main features of successful performance in the vast majority of cases. The profile enables individual manifestations of linguistic material to be viewed within the context of what else is being produced. The same nativelike sentence in two different texts may, in this approach, invite different interpretations on the basis of the profile of the text as a whole.

The coders were provided with detailed guidelines for categorizing texts, by means of font colour, into three basic component types: 'material copied from the question' (coded red), 'non-nativelike material' (coded pink) and 'nativelike material' (coded blue). As outlined below, the last category was sub-divided, and a further category (green) was used as a 'buffer' for unclassifiable text (see later). The coders were instructed to allocate a colour to the first word or words of script, and to continue to allocate that colour until the text no longer fell into that category. In this way, coders focussed on the linguistic coherence of words into strings, rather than judging each word in isolation.

Band: / Centre	3	4	5	6	7	8	Totals
AU i	2	0	0	0	1	1	4
AU ii		1	1	0	2	0	0 4
AU iii	0	1	0	5	1	0	7
AU iv	0	0	0	2	1	0	3
AU subtotal	**3**	**2**	**0**	**9**	**3**	**1**	**18**
FJ i	0	0	0	0	0	1	1
FJ subtotal	**0**	**0**	**0**	**0**	**0**	**1**	**1**
HK i	3	3	6	24	36	11	83
HK ii	2	12	24	0	5	0	43
HK subtotal	**5**	**15**	**30**	**24**	**41**	**11**	**126**
MY i	0	0	10	9	1	0	20
MY ii	0	0	1	3	3	0	7
MY subtotal	**0**	**0**	**11**	**12**	**4**	**0**	**27**
NZ i	0	0	1	0	1	0	2
NZ ii	0	0	0	2	0	0	2
NZ iii	0	1	0	0	0	0	1
NZ iv	0	0	0	1	1	0	2
NZ subtotal	**0**	**1**	**1**	**3**	**2**	**0**	**7**
TW i	1	6	11	4	2	0	24
TW ii	2	8	2	4	4	0	20
TW iii	0	5	0	0	4	1	10
TW subtotal	**3**	**19**	**13**	**8**	**10**	**1**	**54**
Totals	**11**	**37**	**55**	**56**	**60**	**14**	**233**

Table 1: Distribution of scripts by band and centre

4.3.1 Material copied from the question

Material copied from the question is, of course, likely to be nativelike in form. While a candidate does need some command of English to harness such material into use in a writing task, it is, nevertheless, somewhat easy to inflate one's performance by relying on wordstrings provided in the rubric. For this reason, in the assessment of IELTS scripts, material copied from the question is not included. In our profiling, however, it was important to keep a tally of this material, as part of the indication of the overall reliance by the candidate on prefabricated material (whether copied or memorized). It should be noted that native speakers might also quote from the question rubric—it is not inherently wrong to do so. Indeed, it can be a sound aspect of exam technique, because it helps keep the answer focussed.

4.3.2 Non-nativelike material

As noted earlier, it might be feasible to assess a performance simply on the basis of how much or how little non-nativelike material there is. However, this fails to reward a candidate for what has been successfully mastered. Furthermore, it could substantially misrepresent the knowledge level of candidates. This is because the greater one's knowledge of a foreign language, the greater one's capacity to take risks with one's performance (Wray and Fitzpatrick 2008). A very low level learner, in order to perform effectively, may place a great deal of emphasis on memorization, and consequently produce convincingly nativelike output of a restricted type and, thus, relatively few errors. Meanwhile, a higher level learner might eschew memorization in favour of greater self-expression, with the result that more errors are made. Fitzpatrick and Wray (2006) found that intermediate learners of English preferred to choose their own, non-nativelike configurations rather than use nativelike equivalents that they had previously memorized, because their own choices gave them a greater capacity to express their perceptions and identity. Therefore, error coding is most valuable in the context of the larger profile.

When coding errors, a word or wordstring that constituted an error (lexical, grammatical or idiomatic) was coloured pink. Pink asterisks were inserted between words in the text where the error was one of omission. In the subsequent tallies, an asterisk counted as a word.

4.3.3 Nativelike material

As already noted, nativelike material may occur in a text for several reasons, and the profiling aimed to pinpoint differences in its occurrence. To this end, the 'nativelike material' category (all coded blue) was subdivided into three types. The first (blue bold) was 'generic material, which, if memorized, would be useful for most texts of this genre'. Classic examples were discourse markers typical of essays, eg, *There are three reasons for claiming that [sentence];* In summary, I believe that [sentence]. As the designation indicates, it would be a good investment on the part of a learner to memorize a set of such wordstrings, since they could be employed in virtually any discursive writing task, not only in test but in general academic and business writing. Such material is classically used by native speakers to construct an essay, and there is therefore nothing inherently wrong with using it. However, it became clear in the analyses that the balance between the generic nativelike material and other types is of some importance in diagnosis.

The second nativelike subtype (blue italic) was 'topic-generic material, which, if memorized, would be useful for texts of this genre that were on particular typical topics'. Classic examples were lexical phrases and clauses such as *the cost of living* and *all of us need money to live on*. Such material is sufficiently generic that a certain amount, if deliberately memorized, might well be worked into an essay. Nevertheless, some effort would have to go into the learning required for different topics (eg, money, education, environment), so as to ensure an adequate set of phrases and sentences for whatever came up in the test.

Herein lies the crux of the matter. If a candidate has memorized enough such topic-generic material to furnish reliable text for any topic that might be set in the test, does that constitute *excessive* memorization, or effective vocabulary learning? To learn words in an appropriate collocational and colligational environment cannot be considered inappropriate. Again, it is clear that only examining the topic-generic material would not necessarily give a sufficient insight into the performance of the candidate. It is the whole profile that provides a means of interpreting the significance of the quantity of this subtype of material.

The third nativelike subtype (unformatted blue) was 'specific material, which, if memorized, would only be of use for responding to this particular writing task prompt'. Working on the assumption that candidates do not have any means of knowing in advance what the essay title would be, it can be inferred that such material

represents genuine nativelike linguistic knowledge, available on demand. This sort of material is, by definition, usually rather unremarkable: nativelike and idiomatic, but lacking the kind of semantic coherence or functional role that would make it worth specifically memorizing. For instance: '[parents should] train them to be more responsible'.

4.3.4 Buffer material

One additional font colour was used in the coding: green, designated for neutral text, that is, text judged not to contain an error but not classifiable any further. This usually meant that it was not possible to decide whether the word or words really were nativelike choices or not. It is likely that much of this material reflected the candidate's attempt to create novel text apposite to the writing task prompt, using his or her knowledge of individual words and grammatical rules. This category was also used where a single lexical item from the question rubric was used, but it was not clear what else could reasonably have been selected, so it could not be confidently designated 'copying'. Thus, the green text category provides a buffer in coding, to ensure that items difficult to categorize could be set aside, rather than potentially skewing the figures in other categories.

4.4 Example coding and profiling

In order to demonstrate the effect of the profiling, a comparison of two texts is provided here, before the full analysis of the dataset is reported in Section 5. Figures 2 and 3 present the profiles, and Table 2 gives the key to the codes used. In order to accommodate the absence of colour in the printed copy, and to assist the eye in making the comparison, the colour codes have been replaced by grey-scale codes, combined according to their likely motivation. Thus, copied and generic nativelike material are joined under the macro-category 'definitely or probably prefabricated'. Topic-generic and novel material are joined as material 'likely to reflect real learning'. In this way, Figures 2 and 3 can be easily compared, to reveal striking differences in the profiles of the two texts. Each cell in Figures 2 and 3 represents a word in the script. In a string of two or more words coded the same, the first cell contains the code, and the digit in the second indicates the number of words (hence also cells) in the string. It is immediately clear that what the Figure 3 text lacks is any quantity of Top and Nov (both shaded dark grey). That is, the nativelike material in that text is, in almost all cases, either copied from the question or sufficiently generic to have been worth memorizing. In fact, there is no text at all marked Nov. The only material 'likely to reflect real learning' has been coded as 'topic-generic': it could have been memorized from, say, a practice writing task response.

	Code	Meaning	
Definitely or probably prefabricated	Cop	Copied material ('red'): appeared in the essay question	narrative like text
	Gen	Generic material ('blue bold'): would be worth memorizing for most essays	
Likely to reflect real learning	Top	Topic-generic material ('blue italic): would be worth memorizing for clusters of essays on a particular type of topic (eg, the environment; comparison of two education systems)	
	Nov	Novel nativelike material ('unformatted blue'): would only be worth memorizing if you knew the specific essay title in advance	
Unclassified 'buffer' material	Buf	Unclassifiable material ('green'): nativelike but not clearly under the writer's control	
Absence of effective learning or fossilized form	Err	Error ('pink'): a form or lexical choice that was non-nativelike	

Table 2: Key to Figures 2 and 3

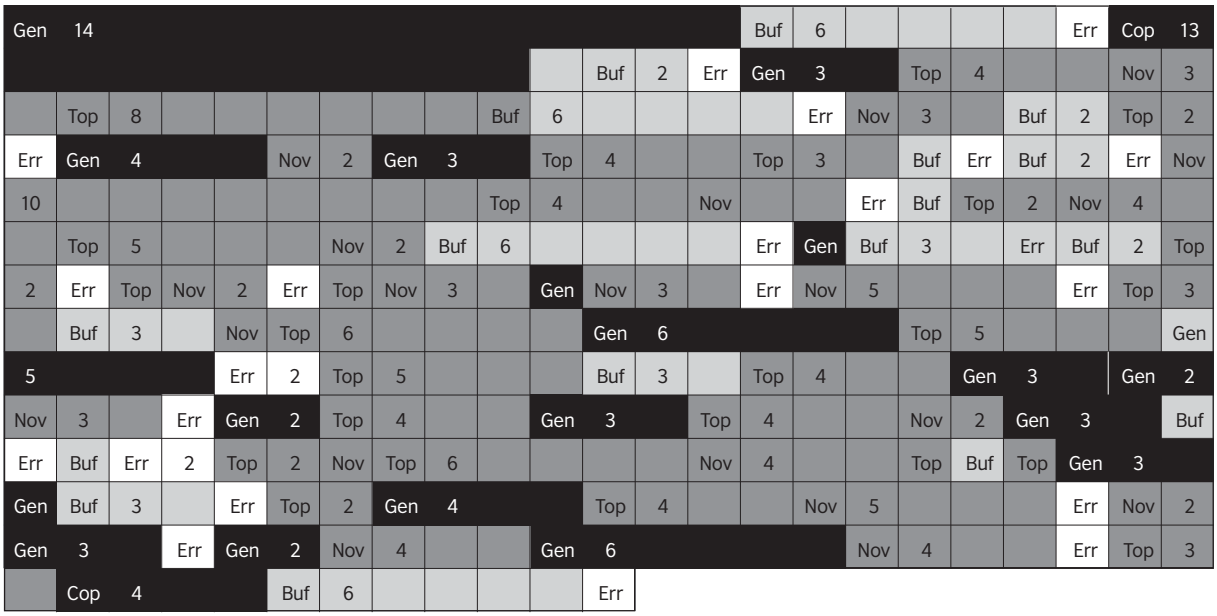


Figure 2: Example of a Band 6 writing task text

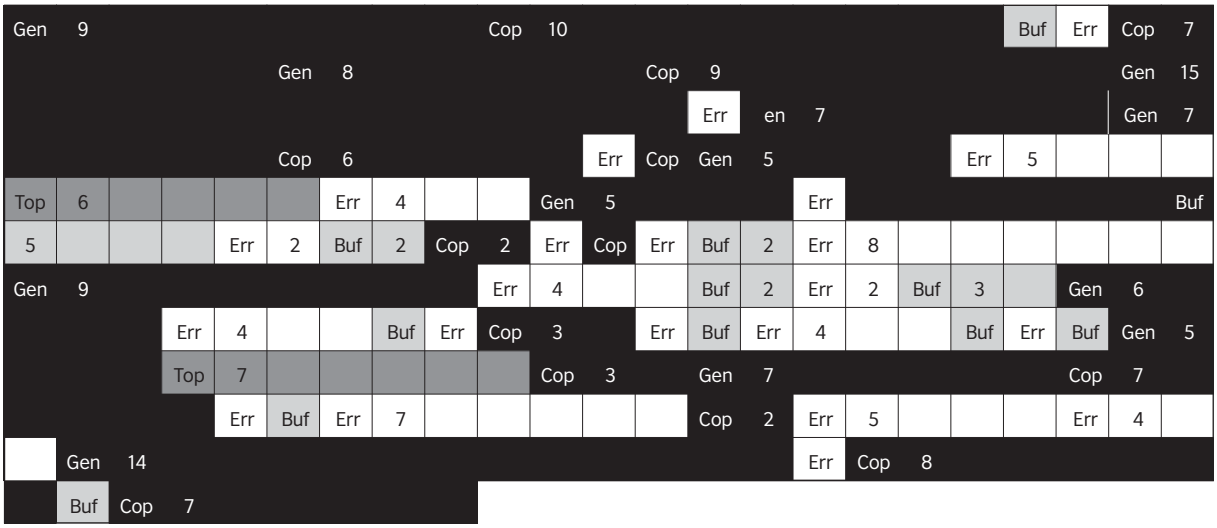


Figure 3: Example of a problematic writing task text

The text in Figure 2 is a typical writing task response given a Band result of 6. We see a strategic use of text copied from the question at the start and end, which can serve to demonstrate relevance. We also see a smattering of generic nativelike text. For instance, the essay begins: "In this essay, I will be presenting my opinion on why I believe that . . .". The second paragraph begins "As we know, . . ." and the final paragraph begins 'In summary, I strongly believe that...'. There are lengthy strings of material classified as topic-generic or novel nativelike material. For instance, in the following extract, the words in italics were classified as topic-generic, and the remainder as novel nativelike: "*in this modern society* money represents everything. *All of us need money to live on*." Errors in this essay are almost all single words, and are usually morphological. In the string "some poorer or less wealthy family", the underlined word was classified as an error, while the preceding words were placed in the 'unclassified' buffer, because they are not sufficiently clearly either nativelike or non-nativelike.

The text represented in Figure 3 was supplied by Cambridge ESOL as particularly problematic in relation to memorized material. It acted as an anchor for our analysis, by helping us identify what sorts of characteristics might be looked for in texts that were somewhat (but less extremely) suspect. It has some very striking features. There is extensive borrowing from the wording of the question, along with a very high reliance on generic nativelike text. For instance, at one point the following occurs: "At first sight, this argument seems reasonable, but if we take a further look, we can find this view can not hold water". The entire string is one that could be used in virtually any discursive essay. ('Can' is underlined to indicate that it was classified as an error. The split of 'cannot' was permitted). Indeed, the first 88 words of the script are either generic material or copied from the question, with the exception of three words classified as errors ('can') just mentioned and two selections of 'good' (to mean 'positive') before 'effect'. In this script, errors were much more likely to be several words in length, indicating problems with structure rather than just morphology or lexis. For example, in these two extracts, the underlined words were classified as errors, while the first two words in the second extract were placed in the buffer category: "computer on the every where" and "email can helps to child fast to give them friends."

We cannot, of course, know how the problematic script came to be produced, nor what the candidate's underlying level of English was: as noted earlier, there is more than one possible explanation for the sequences of nativelike material. However, we do know that this script prompted concerns from at least one IELTS examiner, regarding the likelihood that applying the assessment criteria – which reward positive features – might over-rate the performance, relative to the co-existing evidence of low level proficiency. Figures 2 and 3 illustrate how the oddity of this problematic script can be pinpointed. In the next section, we demonstrate more fully the potential of this profiling to differentiate scripts.

5 THE PROFILE OF 233 IELTS WRITING TASK 2 (ACADEMIC) ESSAYS

As noted earlier, there was one non-significant correlation between the coders. It gave cause for concern regarding the reliability of the coding of topic-generic nativelike and novel nativelike material. Discussions with the coders revealed a lack of confidence about how to differentiate exemplars of these two types, and this extended to their concern about reliability within their own coding of them across scripts. Therefore, these two subtypes were amalgamated in the main profiling analyses. For the reasons already discussed, this was not in fact a particularly problematic compromise to make, since it can be argued that the breadth of memorization necessary for mastering sufficient different topic-generic expressions to cover all possible writing task topics constitutes evidence of genuine learning, rather than inflated, unrepresentative knowledge. The following analyses are focussed around Research Questions 2-6, identified earlier.

5.1 What is the relationship between profile features and band score?

Figure 4 presents an overview of the profiles, using the mean number of words for each text type by band. As would be expected, the amount of non-nativelike (error) material decreases significantly as the band rises (r = -.993, $p < 0.01$). In all of these calculations, Spearman's *rho* has been used, on the basis that bands are based on real scores. There is, however, an argument that the bands are not equally spaced (Ohlrogge 2007), so that a non-parametric test should be used. Pearson's rank correlations result in the same significant correlations as reported here. All probability statements are 2-tailed.). In addition, the amount of non-/topic-generic nativelike material reliably increases by band (r = .996, $p < 0.01$). This, too, is precisely what should happen with reliable banding procedures: the IELTS grading is reflecting the extent to which candidates are capable of expressing apposite content in a nativelike way. However, the tendency to copy material from the question does not significantly correlate with band score (r = -.716)—see Figure 5. This means that the amount of copied material cannot be viewed as indicative of proficiency. Finally, the profiles reveal that there is a significant correlation between the amount of generic nativelike material—usually for organizing the discourse of the essay – and band score (r = -.879, $p < 0.05$) – see Figure 6. However, as the next section will indicate, this is due to one particular test centre.

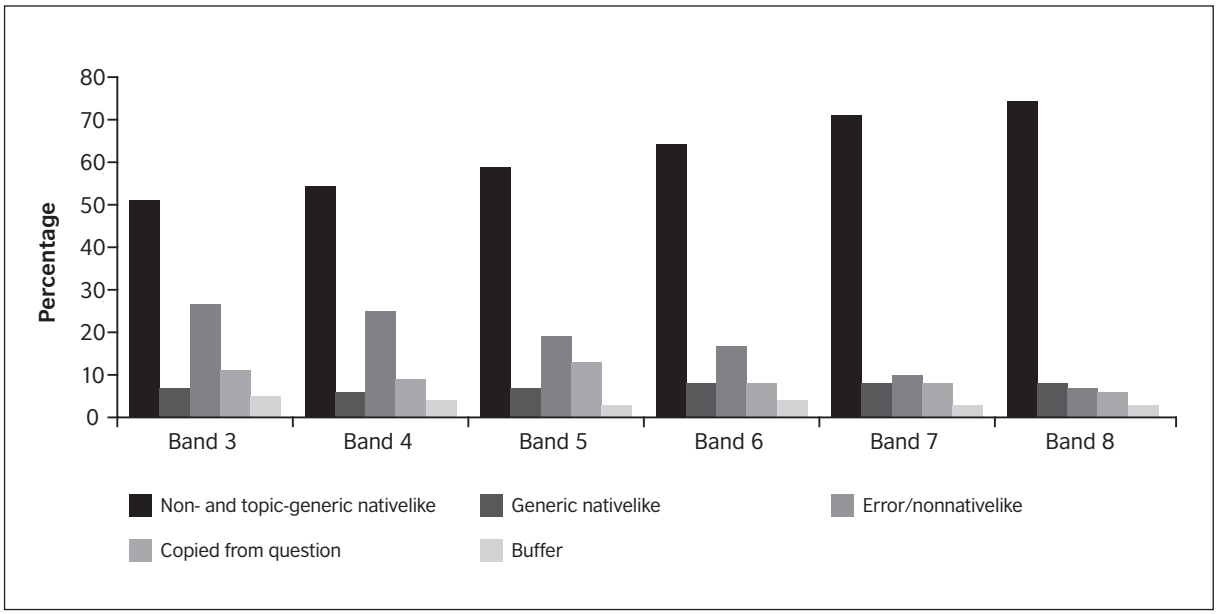


Figure 4: Mean number of words for each text type, by band

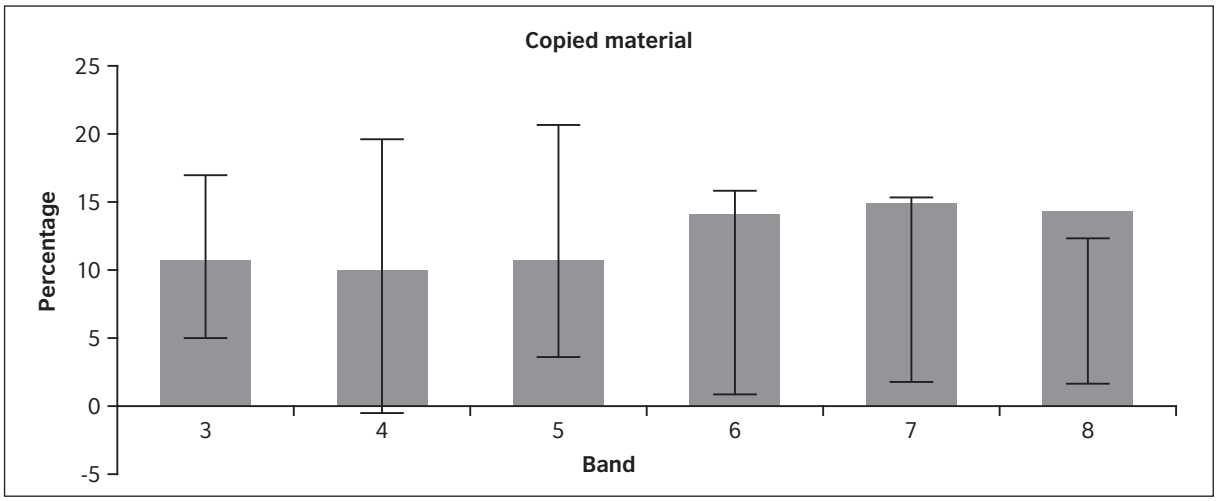


Figure 5: Mean number of words for text copied from the question by band

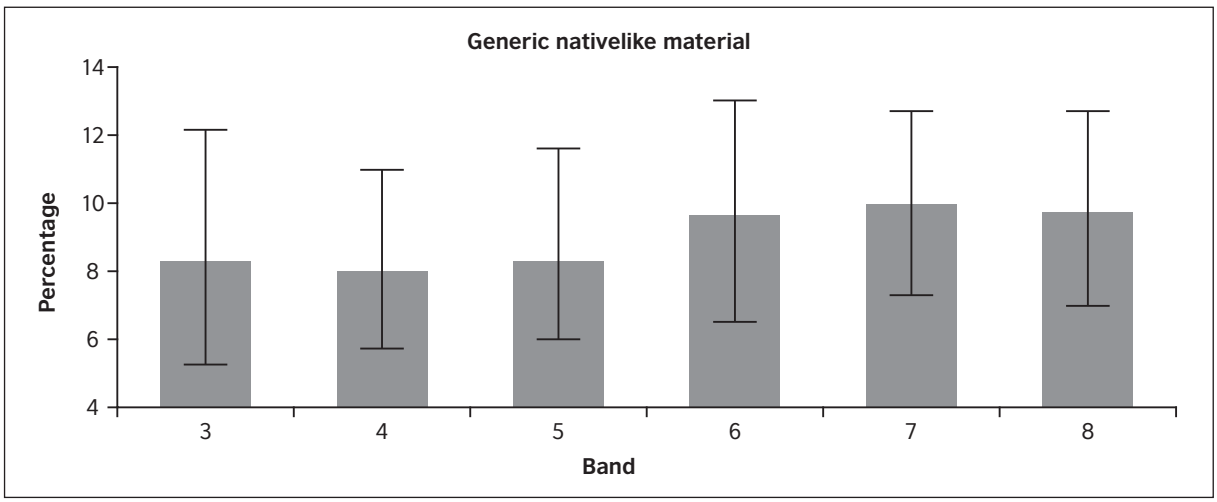


Figure 6: Mean number of words for generic nativelike material by band

5.2 Do candidates from different test centres display different profiles in relation to the amount of potentially memorized material they use?

As noted above, there was a significant correlation between band score and the amount of generic nativelike material. However, an analysis of the profiles by band score indicates that this was due to one centre only. The calculation excluded the Fiji and New Zealand centres since there were too few candidates for reliable figures. While there was a consistent increase in the percentage of generic material in the essays from Hong Kong (r = .985, $p < 0.01$), this was not the case for Australia (r = -.047), Malaysia (r = .583) or Taiwan (r = .575).

A variable of some potential importance in relation to the amount of generic material was the length of the strings so produced. Lengthy strings of memorized material would particularly create the impression of linguistic competence and command of the discourse. Because of the variation in the samples, which could affect the mean, the median lengths were calculated, again excluding the Fiji and New Zealand centres. Hong Kong scripts tended towards longer strings with increased proficiency (r = .892, $p < 0.05$), but the other centres did not: Australia (r = .642), Malaysia (r = .788), Taiwan (r = .495).

These findings may suggest a difference in the style of teaching in Hong Kong, as compared with the other centres. However, the present investigation relies on rather few scripts for some of the other centres used in the comparison (Table 1), and so a larger survey is needed, to establish whether the observed effect is reliable. As to its cause, here too, further research would be advisable. Evidence of a cultural dimension should be explored in the context of comparisons with scripts from some of the 31 IELTS centres in mainland China—no such scripts were available in this study.

5.3 Is it possible, on the basis of norm referencing by profile, to identify problems in a suspect script?

Figure 7 repeats the profiles from Figure 4, but adds the percentage distributions of text types in the problem script described and profiled in Section 4. The distribution is very strikingly different. This indicates that the profiling approach is able both to distinguish a problematic script and to demonstrate why it is problematic.

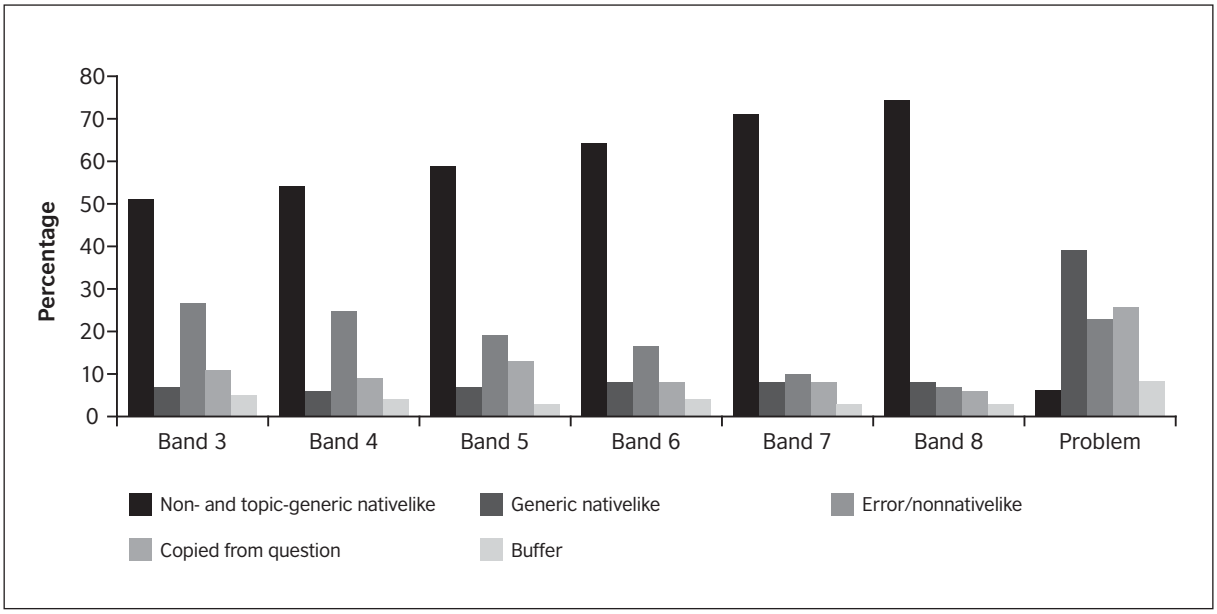


Figure 7: Percentage profiles by band, compared with the problematic script

5.4 What is the simplest profile measure that can be used as the basis of diagnosis?

When an examiner suspects that a script is problematic, it would be possible to adopt the profiling approach described above in order to ascertain the extent to which the script diverges from the norm, and on what basis. However, it would be convenient for such examiners if there were a short cut approach to the same overall discoveries, that could be administered more quickly and with fewer criteria to differentiate. Any such shortcut must gain its credentials from the more detailed procedure as a whole, and so care needs to be taken regarding its identification.

Although the entire profile of the problematic script is at odds with the normal patterns (Figure 7), the shape of that profile is determined by the fact that, in percentage measures, a decrease in one feature entails an increase in another. Although it is the case that in the problem script there was an excessive amount of generic nativelike, copied, and error material, we have seen that only the last of these reliably correlates with proficiency. The tendency in the Hong Kong scripts notwithstanding, both the amount of generic nativelike material and the amount of copied material appear to vary somewhat independently of proficiency, probably for the reasons discussed earlier, namely that they can be indicative of both non-nativelike and nativelike strategies. Therefore, it would be unwise to use either of those aspects of the profile as the focus for a simplified approach to diagnosis.

Using the error measure alone is feasible, but it goes against the spirit of IELTS assessment, which does not focus on errors made but on the extent of the nativelike performance. Applying the error rate aspect of the profile alone would suggest that the problematic script should be graded at Band 4, but giving it such a banding would entail ignoring several key criteria that contribute to the normal profile for that band.

The single most striking feature of the profile of the problem script is the very low proportion of nativelike non-/topic-generic material. This is clearly represented in Figure 8, and it suggests that fast-profiling which homes in on the amount of such material could successfully represent the weaknesses in a problematic script. Specifically, the low percentage of non-/topic-generic nativelike material results from the high percentages of errors, generic and copied material, and represents the extent to which the candidate is able, when not supported by prefabricated material of some kind, to write in error-free English. On the basis of this measure, the script could not be confused with a regular Band 4 script, even though the level of errors is very similar. It should be emphasized that shortcut profiling on the basis of non-/topic-generic nativelike material is *not* the same as simply identifying all the nativelike (or 'correct') language, since generic nativelike text is excluded here.

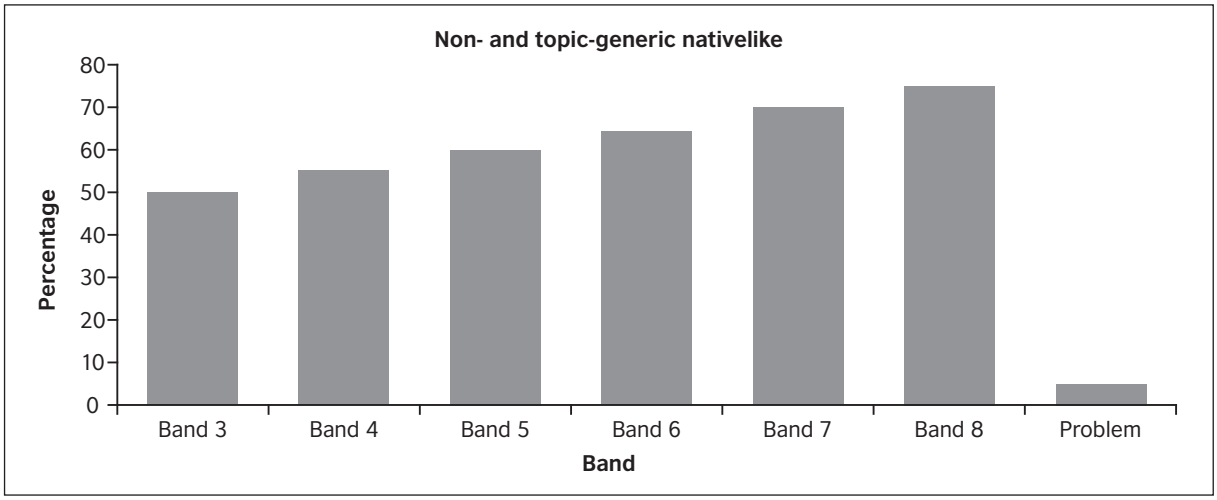


Figure 8: Percentage of non-generic nativelike material in each band and the problematic script

5.5 Is it possible to locate scripts on a continuum, in relation to less striking tendencies towards the overuse of memorized material?

Although in this research all of the scripts in the sample have been used to create the norms, we can reasonably ask whether an examiner could look, within such a sample, for less extreme tendencies towards the kinds of features found in the problem script. In the light of the proposals above, it would be feasible to fast-profile on the basis of the percentage of non-/topic-generic nativelike material and where that is found to be abnormally low, to examine the script(s) for a more general profile.

Table 3 presents the mean, standard deviation and range of percentages of non-/topic-generic nativelike material for each band, along with the thresholds for plus and minus 2.5 standard deviations from the mean. Three scripts in the sample fall below that lower threshold: that is, have an uncharacteristically low percentage of the sort of nativelike text that is likely to reflect true learning.

band	mean	sd	lowest	highest	- 2.5 sd	+ 2.5 sd
3	50.5236	11.1296	34.52	70.83	22.69959	78.34769
4	55.3438	11.5242	**26.01**	81.38	26.53331	84.15426
5	59.334	8.73142	37.85	76.69	37.50545	81.16255
6	64.3977	9.41268	**39.94**	82.19	40.86599	87.92937
7	70.613	8.21158	**48.65**	91.21	50.08406	91.14194
8	74.86071	6.151836	60.13	85.75	59.48112	90.2403

Table 3: Non-/topic-generic nativelike material by band, with +/- 2.5 s.d

The full profile of these three scripts is presented in Figure 9, alongside the norm profile for their respective bands and the original 'problem' script.

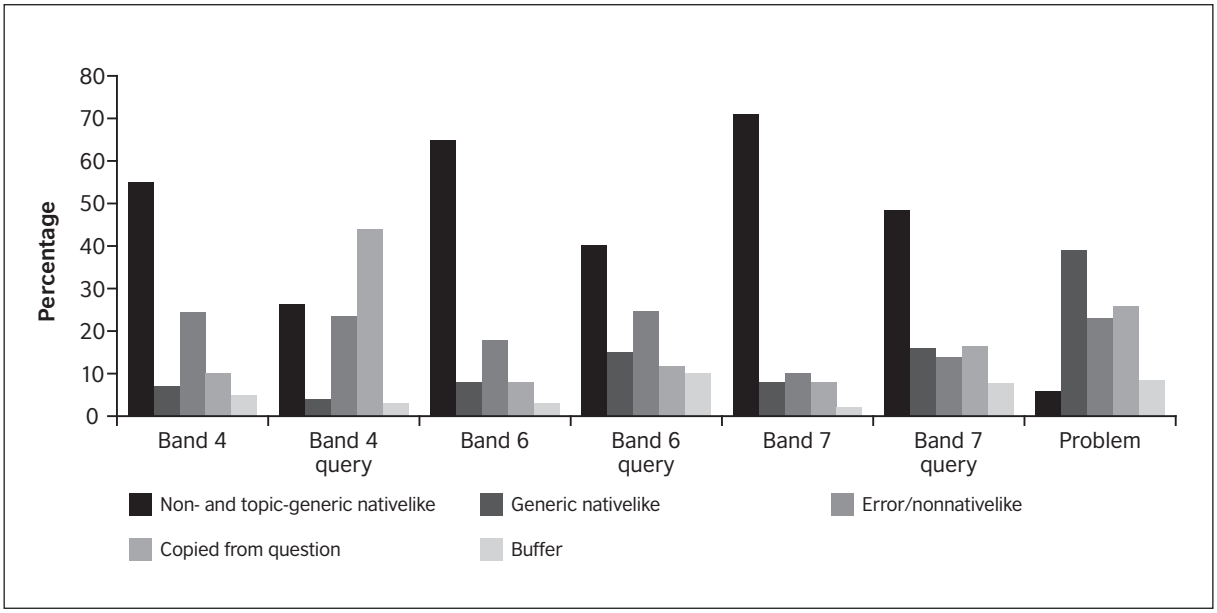


Figure 9: Comparison of queried scripts and their band means

The queried Band 4 script gained its overall assessment outcome on the basis of: Task Response (TR) Band 3, Coherence and Cohesion (CC) Band 5, Lexical Resource (LR) Band 4 and Grammatical Range and Accuracy (GRA) Band 4. It can be seen to differ from the norm specifically in respect of the amount of copied material (4th column). While this *could* indicate that the candidate has been overgraded—in the sense that he or she has not actually provided much evidence of ability—the full profile enables us to see that this script is in two important respects different from the original 'problem' script and similar to the Band 4 norm.

Firstly, the quantity of errors is no higher than the norm. Of course, the less one tries to produce novel nativelike material (as opposed to nativelike material copied from the question) the less one is likely to make errors. However, that maxim did not prevent a high number of errors in the original problem script, so it is notable that the same issue does not arise here. Secondly, in this Band 4 script, there is no inflation in the amount of memorized generic material. This fact could be interpreted as indicating that the candidate did not specifically cram for the test in order to inflate his or her score, so that the banding awarded is indeed representative of the real ability even though it has been derived on the basis of relatively little evidence.

In contrast, the profile of the queried Band 6 script (which gained a run of straight Band 6s on the four assessment criteria subscales) reveals that the low level of non-/topic-generic nativelike material is due to increases, relative to the norm, in errors, generic nativelike material and copied material. This makes the profile a little more similar to that of the original problem script. However, with 40% of the text still non-/topic-generic nativelike, as compared with 5% in the original problem script, the issue is much less extreme. Potentially, the

combined presence of copied and generic material could create an inflated impression of linguistic command relative to the actual evidence for it, since it seems that the candidate is avoiding the production of novel material, and when it is produced it is relatively more likely than the norm to lead to errors.

The Band 7 script (TR 8, CC 7, LR 7, GRA 7) features a lower proportion of nativelike non-/topic-generic material than the norm for even Band 4. However, the level of errors, although above the Band 7 norm, is below the norm for lower bands, indicating that the band assessment is correct. The candidate has relied a little more than normal on both copied and generic material, and, as with the Band 6 script, this could tend to create an impression of a little more ability than is actually the case. However, the profile remains very different from that of the original problem script, and should not raise particular concerns.

Finally, it may be noted that in the queried Band 6 and 7 scripts and problem case (though not the Band 4 one), the amount of unclassified material is also above the norm. As noted earlier the main reasons for material to be unclassified are that it consists of single words from the question (where it would be unreasonable simply to infer 'copying' because it is not clear that a synonym would be appropriate), or because a word or wordstring is neither *non*-nativelike nor very obviously nativelike. It may be that the above-normal level of unclassified material goes some way to explaining the reduced level of non-/topic-generic nativelike material. One explanation could be inconsistency in the coding, and although we do not believe that to be the case here, it would certainly be sensible, for any text under scrutiny, to examine what has been left unclassified, and ascertain whether some of it could in fact be classified.

Another possible explanation is that certain scripts have a specific property, namely, the language is marginal in its nativelikeness. Since the nativelikeness judgement covers form, idiom and lexical choice, what we may be seeing in these scripts is evidence of a particular type of language knowledge, whereby the candidate is, relative to the norm, less capable of recalling idiomatic combinations. Such an individual could have an extensive knowledge of words and grammatical rules, and apply them appropriately, to produce meaningful and formally correct configurations that do not sound nativelike. This approach to production is indeed recognized to be a major feature of adult language learning, and perhaps the single most potent reason why learners fail to attain fully nativelike competence (Wray 2002a).

What this possibility emphasizes is the significance for learners of being able to focus on multiword configurations during learning if they want to produce nativelike output. They would need to target wordstrings in all three of the original profiling subcategories for nativelike material: generic text that can be used in virtually any essay (eg, *To sum up, it is possible to conclude that...*); topic-generic material that can be used for essays on a particular topic or set of topics (eg, *rises in the cost of living; looking after the environment in a time of global warming*); and non-generic material – only worth specifically memorizing if one knows in advance what the writing task prompt is going to be (eg, *children do not respect their parents as much as they used to*). For some learners, generic material, mostly discourse markers, may constitute the bulk of any memorization they do – it furnishes the greatest return for the least effort. More studious learners may be those who are willing to learn the *topic*-generic material that tends to be encountered when reading around and writing about the kinds of topics that typically come up in the test. Meanwhile, the really successful learners – those who are on the most promising trajectory towards high level proficiency – may be the ones who are capable of, and committed to, internalizing nativelike nuances as a matter of course. For instance, Ding (2007) notes of one of his informants that 'while other students used 'Family is very important', she borrowed a sentence pattern she had learned from [a textbook]: 'Nothing can be compared with the importance of the family'. This made a better sentence, she said' (p.277).

While memorization remains, by definition, a relatively unpalatable and impractical solution for most learners in respect of the most open subcategory, non-generic nativelike material, research suggests that extensive memorization has additional benefits for learning over simply providing access to that particular material in the future (Ding 2007; Qi 2006; Ting and Qi 2001). It opens the door to a 'feel' for the language, and instils confidence.

6 CONCLUSION

Although even one of Ding's (2007) ultimately successful learners regarded memorization, in the early stages of his learning, as "the most stupid method in the world" (p 278), there is clear evidence that "with repeated practice... [an] initially noticed new feature becomes familiar and is transferred from the working memory to the long-term memory, retrievable when need arises" (Ding 2007, p 279). Such transfer can lie at the heart of truly successful learning, and so it is important that testing does not treat it with undue suspicion. Furthermore,

memorizing sufficient linguistic material to create a plausible product in test conditions entails a great deal of work, so that viewing it simply as a form of 'cheating' would be inappropriate.

This report has demonstrated how it may be possible to establish, for both extreme and borderline scripts, the basis of an examiner's disquiet. In essence, a rough estimation is made of the amount of nativelike material that it is reasonable to infer reflects true knowledge: appropriately used non- and topic-generic nativelike language. The rationale is that the candidate must either have constructed it from scratch or else have retrieved it from such a large store of memorized material that, by virtue of its availability, it must be credited as the product of real learning.

6.1 Recommendations to IELTS examiners

This study is able to make some first recommendations for the future training of IELTS examiners, regarding scripts that appear to have excessive memorized material. Firstly, the strong negative correlation in our sample between band score and the level of errors, and the positive correlation between band score and non-/topic-generic nativelike language, indicate that the banding procedures are robust. The reason why potentially memorized material is problematic is precisely because it does not correlate with proficiency (though it did for Hong Kong scripts—see earlier). This means that examiners should have confidence in their intuitions regarding suspicious scripts.

Secondly, the evidence that memorization can be the path to effective learning, in both a first and second language, coupled with the fact that native speakers legitimately internalize useful turns of phrase as part of their own preparation for tests and exams, jointly create a dilemma in relation to whether it is appropriate to reward apparently memorized material. Examiners need not, therefore, feel that it is up to them to *solve* the problem of a suspect script: the difficulty is inherent and essentially insoluble, since there is no independent way to tell what the candidate truly 'knows' (nor any uncontentious way to define 'knowledge' in this regard).

Thirdly, if faced with a perplexing script, the examiner can adopt the simplified profiling approach described in this report, by highlighting continuous runs of linguistic material falling into the category 'nativelike non- or topic-generic': that is, material that is nativelike but not copied from the question, and that would not be worth memorizing for generic use across all written tasks. Isolated words, ie, words that are surrounded by material that would fall into another category, should not be counted as non-/topic-generic (see earlier description of buffer material). It is recommended that the procedure be carried out not only for the suspect script but also for a handful of uncontentious others, as a means of gauging the reliability of the coding relative to the norms provided here.

By counting the total number of words highlighted, and comparing them to the norm for the band the script appears to fall into, it should be possible to ascertain whether there are grounds for identifying the suspect script as abnormal (and the others profiled at the same time as normal). The present study suggests the following norms (based on a lower threshold of 2.5 standard deviations from the mean), though further research should be done on much larger samples to confirm these values. In particular, the Band 5 threshold, as determined in this study, seems possibly a little high relative to the others.

Band 3 scripts contain no less than 22% non-/topic-generic nativelike material.

Band 4 scripts contain no less than 26% non-/topic-generic nativelike material.

Band 5 scripts contain no less than 37% non-/topic-generic nativelike material.

Band 6 scripts contain no less than 40% non-/topic-generic nativelike material.

Band 7 scripts contain no less than 50% non-/topic-generic nativelike material.

Band 8 scripts contain no less than 59% non-/topic-generic nativelike material.

As the analyses in Section 5 showed, scripts falling below the threshold are not necessarily irregular—something that can be ascertained by examining other features of the profile. A truly problematic script, such as the one profiled in Section 4, will be strikingly different in regard to the distribution of the profile components.

The purpose of the profiling should never be construed as that of 'proving' that material in a script has been memorized. That simply is not possible. Rather, profiling offers a means by which the examiner can offer a justification for his/her disquiet, as part of the case for a review of the script.

6.2 Recommendation to IELTS

We have argued in this report that a certain amount of memorized material in a script is not only acceptable but an indicator of task proficiency. We have also shown that in a given sample, such as the 233 scripts

examined here, probably none at all will raise real concerns of the kind associated with the problem script analyzed in Section 4. The normal band profiles amply demonstrate that IELTS examiners are well-trained in recognizing and rewarding a healthy balance of novel and potentially memorized material, and that the criteria are well constructed to enable it. There is only a problem when examiners are confronted with a script in which the sheer quantity of possibly memorized material threatens to distort the score.

Our first recommendation to IELTS regards raising examiners' awareness of both the potential impact of excessive memorization on a script, and the ways in which a script can be profiled to assist in identifying the problem.

Our second recommendation is that some consideration be given to the main reason, as we perceive it, why a problem script could appear to justify a higher band score than the examiner feels it truly deserves. This reason takes us back to the theoretical underpinning of research into formulaic language.

When a person constructs novel language from scratch, three types of knowledge are required: what it is appropriate to say, which vocabulary to select, and how to arrange it grammatically. These knowledge types correspond to three of the four components of the IELTS banding: Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Novel material, therefore, legitimately deserves a reward in relation to each of these three components. Memorized material, however, compromises the independence of the components. It must still be appropriately used within the text, so Coherence and Cohesion should be rewarded. However, the candidate's demonstration of Lexical Resource does not include the individual selection of each word, only the selection of the complete sequence, drawn from the mental lexicon like a single unit (Wray 2002a). That is, the wordstring's lexis is pre-specified. In the same way, although the wordstring must be correctly embedded grammatically into the surrounding text, no specific decisions need to be made regarding the grammatical forms, since they too are pre-specified. The crux of the matter, then, is that if a memorized wordstring is treated like a novel wordstring, it could be rewarded on the basis of lexical selections and grammatical decisions that were not made.

One solution to this conundrum would be to view the wordstring as a single lexical choice. Its selection as a *single item* can then be rewarded either under Coherence and Cohesion (if it is appropriately used to structure the discourse) or under Lexical Resource (if it is a content expression), without rewarding its individual components—in the same way as one might reward, as a single item, the use of the French expression in 'he displayed a certain *je ne sais quoi*' without attributing to the user the capacity to create novel sentences in French. Its grammatical place within the text, also, could be rewarded, under Grammatical Range and Accuracy, on the same basis as the correct grammatical use of a single word, without rewarding the internal grammatical configuration—just as one might reward the correct grammatical embedding of the idiom 'if I were you' without assuming that the writer had a full command of the subjunctive. By regarding generic multiword strings as single vocabulary items, it would be possible to reward the use of a broader than average range of them in the same way as one rewarded a broad single word vocabulary: learners typically internalize a few discourse markers and overuse them (Granger 1998). Similarly, as with single words, they could be rewarded for being register-appropriate. Swedish learners have been found to use inappropriately informal multiword strings in written contexts (Wiktorsson 2003).

Key theoretical considerations impact on the practicality of treating potentially memorized wordstrings like single words for assessment purposes. Firstly, there is the question of identifying what counts as potentially memorized. In this research we have allowed the analyst (and, through future wider implementation, also the examiner) to make the judgement intuitively. Our sense is that even in the context of assessment, that approach remains the most appropriate. Examiners need to feel empowered to draw on both their knowledge of the language and their experience in the examining role, to sense the likelihood that a given wordstring has been memorized and—importantly—to evaluate the impact, positive or negative, of its inclusion. Training can support the development of examiners' confidence in this regard.

The second theoretical issue regards the fact that one does not always memorize a *complete* string. The most useful wordstring to memorize might be one with gaps in it, such as 'The most important issue with regard to -- ___ is ___' and 'several issues can be identified. Firstly ___. Secondly ___. Thirdly ___. [etc.]' Clearly one needs to treat the unchanging frame as a single word, but reward the varying items within it as independent choices.

The third theoretical issue regards the fact that memorization is not always perfect (Fitzpatrick and Wray 2006; Wray and Fitzpatrick 2008). This means that attempts to reproduce memorized wordstrings may contain errors. They would be dealt with in the same way as morphological or spelling errors within a single word though, of course, rather more errors could accumulate in a wordstring.

Thus it can be seen that we are not, here, by any means suggesting that the assessment criteria used by IELTS examiners be changed. On the contrary, since it is only extreme cases of memorization that are problematic, doing so really would be using a sledgehammer to crack a nut. Rather, we have drawn attention to various issues relating to the assessment of productive skills in writing—ones that affect all tests and agencies—and indicated ways in which IELTS examiner training can introduce a practical approach to their resolution when they arise.

6.3 Future research

The aims of this research were to investigate the effect of memorization on the writing test scripts (Academic Task 2) of Chinese mother tongue IELTS candidates, to develop a tool for profiling scripts in this regard, and to streamline the tool for easy use by any examiner, in order to help pinpoint the basis of disquiet about a script. This particular mother tongue group was selected because of the historical and well-documented strategy of using memorization as a learning tool in China. Because the tool was designed for use by examiners, we specifically did not develop a software-based diagnostic, nor one that relies on statistical analyses carried out by the examiner. Nevertheless, there is, of course, scope to develop the profiling tool in that direction. Since there is an inherent weakness in any profiling technique that relies on intuitive judgements, one possibility for the future is to replace this element with automatic profiling based on separate sweeps for different feature types, and, probably, referring to an extensive lexicon of generic discourse marker phrases. In the meantime, the results of the present study could usefully be confirmed through an extended replication. Again, given the vulnerability of hand-coding, a priority should be the verification of the robustness of the coding procedures and a full validation of inter-coder reliability. As noted earlier, it would also be informative to explore the reasons for the correlation between band score and both the overall amount of generic nativelike material and the mean length of continuous strings of it in the Hong Kong scripts. Ideally, a larger study might be undertaken, not only comparing Hong Kong scripts with those from other centres inside and outside mainland China, but also exploring more qualitatively, through observation, the methods by which teachers prepare students for the IELTS test in different places.

This research contributes to the body of recent work on the role that formulaic sequences play in the construction of discourse in tests by second language learners (eg, Ohlrogge 2007; Read and Nation 2006). However, key to making decisions about how to assess such material is understanding the processes by which it is available for production, and the complex reasons why it is used. A learner's use of material that *could* be memorized does not mean that it *was* memorized. Furthermore, if it *has* been memorized, its use could be indicative of low or high proficiency. The enigma is that it is simultaneously eminently nativelike and eminently *non*-nativelike to use certain kinds of common linguistic expressions correctly. There is, in consequence, no way of judging formulaic language without reference to the rest of the linguistic profile.

REFERENCES

Au, C and Entwistle, N, 1999, 'Memorization with understanding in approaches to studying: cultural variant or response to assessment demands', paper presented at the European Association on Learning and Instruction Conference, Gothenburg, August.

Cooper, B J, 2004, 'The enigma of the Chinese learner', *Accounting Education* vol 13, pp 289-310

Dahlin, B and D Watkins, 2000, 'The role of repetition in the processes of memorizing and understanding: A comparison of the views of German and Chinese secondary school students in Hong Kong', *British Journal of Educational Psychology* 70, pp 65-84

Ding, Y, 2007, 'Text memorization and imitation: the practices of successful Chinese learners of English', *System* vol 35, pp 271-280

Fitzpatrick, T and Wray, A, 2006, 'Breaking up is not so hard to do: individual differences in L2 utterances in L2 utterance memorization', *Canadian Modern Language Review* vol 63, no 1, pp 35-57

Granger, S, 1998, 'Prefabricated patterns in advanced EFL writing: collocations and formulae', in *Phraseology: theory, analysis and applications*, ed. A P Cowie, Clarendon Press, Oxford, pp 145-160

Ho, I, Salili, F, Biggs, J and Hau KT, 1999, 'The relationship among causal attributions, learning strategies and level of achievement: a Hong Kong case study', *Asia Pacific Journal of Education* vol 19, no 1, pp 44-58

Kennedy, P, 2002, 'Learning cultures and learning styles: myth-understanding about adult (Hong Kong) Chinese learners', *International Journal of Lifelong Education* vol 21, no 5, pp 430-445

Marton, F, Dall'Alba, G and Tse, L K, 1993, 'The paradox of the Chinese learner', Occasional Paper no 93.1, RMIT, Educational Research and Development Unit, Melbourne

Ohlrogge, A, 2007, 'Deceptively memorized or appropriately stored whole? The use of formulaic expression in intermediate EFL writing assessment', paper presented at the Formulaic Language Symposium, University of Wisconsin, Milwaukee, April 18-21

Qi, Y, 2006, *A longitudinal study on the use of formulaic sequences in monologues of Chinese tertiary-level EFL learners*, Unpublished PhD thesis, School of Foreign Studies, Nanjing University

Read, J. and Nation, P, 2006, 'An investigation of the lexical dimension of the IELTS Speaking Test', *IELTS Research Reports* vol 6, pp 207-231

Ting, Y, and Qi, Y, 2001, 'Learning English texts by heard in a Chinese university: a traditional literacy practice in a modern setting', *Foreign Language Circles* vol 5, pp 58-65

Wiktorsson, M, 2003, *Learning idiomaticity*, Lund Studies in English 105, Lund University, Sweden

Wray, A, 1999, 'Formulaic language in learners and native speakers' *Language Teaching* vol 32, no 4, pp 213-231

Wray, A., 2000, 'Formulaic sequences in second language teaching: principle and practice', *Applied Linguistics* vol 21, no 4, pp 463-489

Wray, A, 2002a, *Formulaic language and the lexicon*, Cambridge University Press, Cambridge

Wray, A, 2002b, 'Formulaic language in computer-supported communication: theory meets reality', *Language Awareness* vol 11, no 2, pp 114-131

Wray, A, 2004, 'Here's one I prepared earlier: formulaic language learning on television', in *Formulaic sequences: acquisition, processing and use*, ed N Schmitt, John Benjamins, Amsterdam, pp 249-268

Wray, A, Cox S, Lincoln, M and Tryggvason, J, 2004, 'A formulaic approach to translation at the Post Office: reading the signs', *Language and Communication* vol 24, no 1, pp 59-75

Wray, A and Fitzpatrick, T, 2008, 'Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence' in *Phraseology in foreign language learning and teaching*, eds F Meunier and S Granger, John Benjamins, Amsterdam, pp 123-148

Wray, A and Staczek, J, 2005, 'One word or two? Psycholinguistic and sociolinguistic interpretations of meaning in a court case', *International Journal of Speech, Language and the Law* vol 12, no 1, pp 1-18

Zhanrong, L, 2002, 'Learning strategies of Chinese EFL learners: review of studies in China' *RTVU ELT Express*, http://www1.openedu.com.cn/elt/2/4.htm [last accessed 17th Feb 2008].

APPENDIX 1: DETAILS OF WRITING TEST

>>> Writing

Duration and format

The Writing test takes 60 minutes. There are two tasks to complete. It is suggested that about 20 minutes is spent on Task 1 which requires candidates to write at least 150 words. Task 2 requires at least 250 words and should take about 40 minutes.

Candidates may write on the question paper but this cannot be taken from the examination room and will not be seen by the examiner.

Answers must be given on the answer sheet and must be written in full. Notes or bullet points in whole or in part are not acceptable as answers.

Task types

Academic Writing

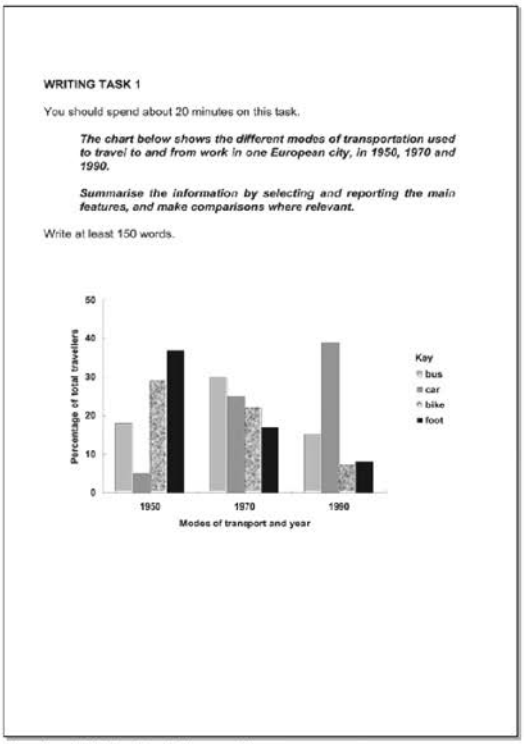
In Task 1 candidates are asked to describe some information (graph/table/chart/diagram), and to present the description in their own words. Depending on the type of input and the task suggested, candidates are assessed on their ability to:

- organise, present and possibly compare data
- describe the stages of a process or procedure
- describe an object or event or sequence of events
- explain how something works

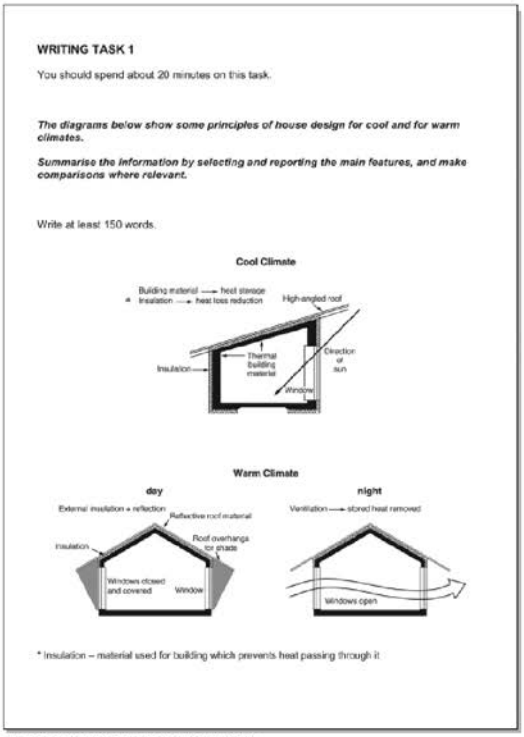
In Task 2 candidates are presented with a point of view or argument or problem. Candidates are assessed on their ability to:

- present the solution to a problem
- present and justify an opinion
- compare and contrast evidence, opinions and implications
- evaluate and challenge ideas, evidence or an argument

The issues raised are of general interest to, suitable for and easily understood by candidates entering undergraduate or postgraduate studies or seeking professional registration.



Academic Writing Task 1 (example)



Academic Writing Task 1 (example)

8 | IELTS Handbook 2007

Marking and assessment
Each task is assessed independently. The assessment of Task 2 carries more weight in marking than Task 1.

Writing responses are assessed by certificated IELTS examiners. All IELTS examiners hold relevant teaching qualifications and are recruited as examiners by the test centres and approved by British Council or IDP: IELTS Australia.

Detailed performance descriptors have been developed which describe written performance at the nine IELTS bands. Public versions of these descriptors are available on the IELTS website.

The descriptors apply to both the Academic and General Training Modules and are based on the following criteria.

Task 1 responses are assessed on:

- Task Achievement
- Coherence and Cohesion
- Lexical Resource
- Grammatical Range and Accuracy

Task 2 responses are assessed on:

- Task Response
- Coherence and Cohesion
- Lexical Resource
- Grammatical Range and Accuracy

Task 1

Task Achievement
This criterion assesses how appropriately, accurately and relevantly the response fulfils the requirements set out in the task, using the minimum of 150 words.

Academic Writing Task 1 is a writing task which has a defined input and a largely predictable output. It is basically an information-transfer task which relates narrowly to the factual content of an input diagram and not to speculated explanations that lie outside the given data.

General Training Writing Task 1 is also a writing task with a largely predictable output in that each task sets out the context and purpose of the letter and the functions the candidate should cover in order to achieve this purpose.

Coherence and Cohesion
This criterion is concerned with the overall clarity and fluency of the message: how the response organises and links information, ideas and language. Coherence refers to the linking of ideas through logical sequencing. Cohesion refers to the varied and appropriate use of cohesive devices (for example, logical connectors, pronouns and conjunctions) to assist in making the conceptual and referential relationships between and within sentences clear.

Lexical Resource
This criterion refers to the range of vocabulary the candidate has used and the accuracy and appropriacy of that use in terms of the specific task.

Grammatical Range and Accuracy
This criterion refers to the range and accurate use of the candidate's grammatical resource as manifested in the candidate's writing at the sentence level.

Task 2

Task Response
In both Academic and General Training Modules Task 2 requires the candidates to formulate and develop a position in relation to a given prompt in the form of a question or statement. Ideas should be supported by evidence, and examples may be drawn from the candidates' own experience. Responses must be at least 250 words in length.

Scripts under the required minimum word limit will be penalised.

Scores are reported in whole and half bands.

APPENDIX 2: SPECIFIC INSTRUCTIONS FOR THE WRITING TASK 2 TO WHICH STUDY PARTICIPANTS RESPONDED (OTHER THAN THE 'PROBLEMATIC SCRIPT')

WRITING TASK 2

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic:

Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

To what extent do you agree or disagree with this opinion?

You should use your own ideas, knowledge and experience and support your arguments with examples and relevant evidence.

Write at least 250 words.

Band	Task response	Coherence and cohesion	Lexical resource	Grammatical range and accuracy
9	fully addresses all parts of the task presents a fully developed position in answer to the question with relevant, fully extended and well supported ideas	uses cohesion in such a way that it attracts no attention skilfully manages paragraphing	uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'	uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
8	sufficiently addresses all parts of the task presents a well-developed response to the question with relevant, extended and supported ideas	sequences information and ideas logically manages all aspects of cohesion well uses paragraphing sufficiently and appropriately	uses a wide range of vocabulary fluently and flexibly to convey precise meanings skilfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation produces rare errors in spelling and/or word formation	uses a wide range of structures the majority of sentences are error-free makes only very occasional errors or inappropriacies
7	addresses all parts of the task presents a clear position throughout the response presents, extends and supports main ideas, but there may be a tendency to over-generalise and/or supporting ideas may lack focus	logically organises information and ideas; there is clear progression throughout uses a range of cohesive devices appropriately although there may be some under-/over-use presents a clear central topic within each paragraph	uses a sufficient range of vocabulary to allow some flexibility and precision uses less common lexical items with some awareness of style and collocation may produce occasional errors in word choice, spelling and/or word formation	uses a variety of complex structures produces frequent error-free sentences has good control of grammar and punctuation but may make a few errors
6	addresses all parts of the task although some parts may be more fully covered than others presents a relevant position although the conclusions may become unclear or repetitive presents relevant main ideas but some may be inadequately developed/unclear	arranges information and ideas coherently and there is a clear overall progression uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical may not always use referencing clearly or appropriately uses paragraphing, but not always logically	uses an adequate range of vocabulary for the task attempts to use less common vocabulary but with some inaccuracy makes some errors in spelling and/or word formation, but they do not impede communication	uses a mix of simple and complex sentence forms makes some errors in grammar and punctuation but they rarely reduce communication
5	addresses the task only partially; the format may be inappropriate in places expresses a position but the development is not always clear and there may be no conclusions drawn presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail	presents information with some organisation but there may be a lack of overall progression makes inadequate, inaccurate or over-use of cohesive devices may be repetitive because of lack of referencing and substitution may not write in paragraphs, or paragraphing may be inadequate	uses a limited range of vocabulary, but this is minimally adequate for the task may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader	uses only a limited range of structures attempts complex sentences but these tend to be less accurate than simple sentences may make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader
4	responds to the task only in a minimal way or the answer is tangential; the format may be inappropriate presents a position but this is unclear presents some main ideas but these are difficult to identify and may be repetitive, irrelevant or not well supported	presents information and ideas but these are not arranged coherently and there is no clear progression in the response uses some basic cohesive devices but these may be inaccurate or repetitive may not write in paragraphs or their use may be confusing	uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task has limited control of word formation and/or spelling; errors may cause strain for the reader	uses only a very limited range of structures with only rare use of subordinate clauses some structures are accurate but errors predominate, and punctuation is often faulty
3	does not adequately address any part of the task does not express a clear position presents few ideas, which are largely undeveloped or irrelevant	does not organise ideas logically may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas	uses only a very limited range of words and expressions with very limited control of word formation and/or spelling errors may severely distort the message	attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
2	barely responds to the task does not express a position may attempt to present one or two ideas but there is no development	has very little control of organisational features	uses an extremely limited range of vocabulary; essentially no control of word formation and/or spelling	cannot use sentence forms except in memorised phrases
1	answer is completely unrelated to the task	fails to communicate any message	can only use a few isolated words	cannot use sentence forms at all
0	does not attend does not attempt the task in any way writes a totally memorised response			